Statistique et Apprentissage Automatique en Science des Données

Partie 2 : Inférer à partir de données observées

Laurent Risser (CNRS) — Année 2026/2027

Partie 2.1 : Introduction

LAROUSSE

inférer

verbe transitif

(latin *inferre,* alléguer)

Conjugaison

Tirer une conséquence de quelque chose, conclure, induire quelque chose de quelque chose : Qu'avez-vous inféré de tels propos ?

SYNONYMES :

conclure - déduire - induire

D'un point de vue plus technique

- On peut souhaiter mettre en lien une variable de sortie Y à prédire à $p \geq 1$ variables d'entrée $X = \left(X^1, X^2, \ldots, X^p\right)$.
- Afin de formaliser cette relation, on utilisera typiquement une fonction h de paramètres θ telle que :

$$\hat{Y} = h_\theta\left(X^1, X^2, \ldots, X^p\right)$$

où $\hat{Y}$ estime la valeur de Y.

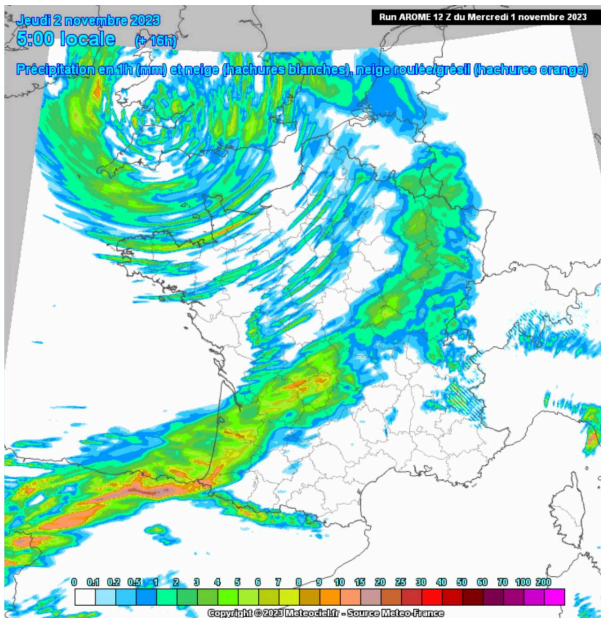
Exemples

- Quelle température fera-t-il demain à Toulouse en fonction de l'état de l'atmosphère aujourd'hui ?
- Quel est le niveau de risque d'un dossier pour un prêt immobilier en fonction du dossier client ?
- Quelle est la contrainte subie par une aile d'avion en fonction des variables de vol ?
- Est-ce qu'une voiture doit freiner en urgence en fonction des informations remontées par des caméras optiques et des capteurs Lidar ?

Il convient de distinguer les modèles *déterministes* et *statistiques* dans la définition de h_θ :

- *Modèles déterministes* : Equation ou ensemble d'équations qui émanent souvent de lois physiques, chimiques, économiques, … et représentent le comportement attendu du phénomène
- *Modèles statistiques* : Il peut être difficile de développer un modèle théorique car le phénomène étudié est trop complexe. On a alors recours à une modèle statistique basé non pas sur une théorie, mais sur des données observées. On parle de **modèle inférentiel**.

Pile ou face ?



Prévisions de précipitation (tempête Ciarán)

Questions au coeur de ce cours :

- Comment modéliser un problème de statistique inférencielle ?
- Comment mettre en lien des entrées et sorties ?
- Quels sont les fondements de l'apprentissage automatique ?
- Résoudre un problème en pratique avec le modèle linéaire ?
- …

Objectif :

- Vous donner les clés pour utiliser les outils d'apprentissage de manière pertinente.

Pour les SDD :

- Approfondissement des notions avec beaucoup d'exercices pour les assimiler

Partie 2.2 : Rappels en Probabilité et Statistique

Expérience

- Chaque étudiant de la classe tire $n = 10$ fois une pièce à pile ou face avec et compte le nombre de fois que la pièce est tombée sur pile. Pile correspond alors à $X_i = 1$ et face à $X_i = 0$.

$\times 10$ → Nombre de « pile » ?

- On suppose que $\mathbb{P}(X = 1) = 0.5$ et $\mathbb{P}(X = 0) = 0.5$, ce qui est sans doute très proche de la réalité. Ainsi l'espérance (moyenne) de X est $m = 0.5$ et son écart type est $s = 0.5$.
- On va dessiner un graphique dans lequel l'abscisse représente le nombre de 'piles' potentiellement obtenus par un étudiant (entre 0 et 10) et l'ordonnée représente le nombre d'étudiant qui ont obtenus ce nombre de 'piles' divisé par le nombre total d'étudiants.
- On constatera que cette courbe approche la densité de la loi normale de moyenne $10m$ et d'écart type $s\sqrt{10}$.

Variable aléatoire : Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera :

- *cas discret*, par exemple si $X \in \{0,1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face (ou client Homme/Femme, patient jeune/adulte/âgé, …).
- *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température (ou age d'un client, taille d'un patient, …).

Loi de probabilité discrète Par exemple si l'on joue à pile ou face avec une pièce parfaitement équilibrée, on a $\mathbb{P}(X = 0) = 1 - p = 0.5$ et $\mathbb{P}(X = 1) = p = 0.5$. On remarquera que la somme des probabilités de tous les résultats possibles dans le cas discret est toujours 1.

Variable aléatoire : Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera :

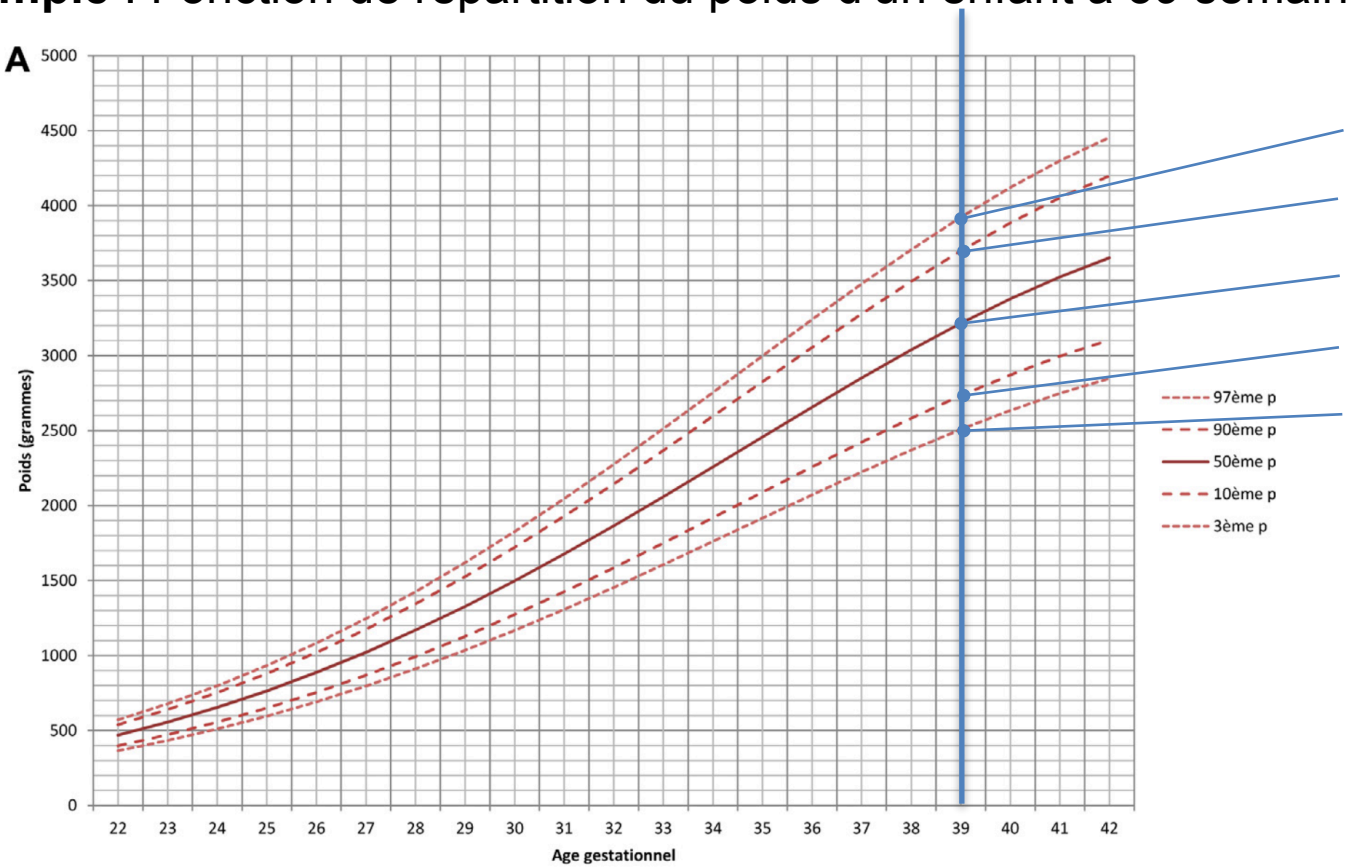
- *cas discret*, par exemple si $X \in \{0,1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face (ou client Homme/Femme, patient jeune/adulte/âgé, …).
- *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température (ou age d'un client, taille d'un patient, …).

Loi de probabilité continue Dans le cas continu, écrire $\mathbb{P}(X = x)$ n'a aucun sens puisque la probabilité d'une valeur exacte est infinitésimale. On pourra par contre utiliser la *fonction de répartition* $F_X(x) = \mathbb{P}(X \leq x)$ pour représenter comment se répartissent les probabilités des différents résultats de X. Il sera alors possible de quantifier les chances que X soit sur une certaine gamme de valeurs $\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Naturellement, on aura toujours $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$.

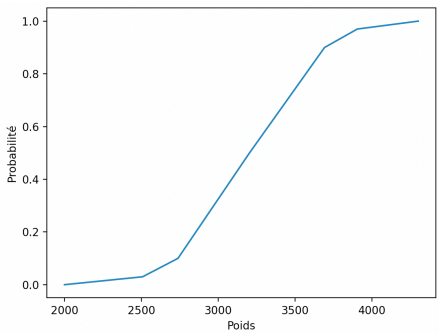
Variable aléatoire : Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera :

- *cas discret*, par exemple si $X \in \{0,1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face (ou client Homme/Femme, patient jeune/adulte/âgé, …).
- *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température (ou age d'un client, taille d'un patient, …).

Exemple : Fonction de répartition du poids d'un enfant à 39 semaines $F_X(x) = \mathbb{P}(X < x)$



$$F_X(3904) = 0.97$$
$$F_X(3691) = 0.90$$
$$F_X(3203) = 0.50$$
$$F_X(2738) = 0.10$$
$$F_X(2508) = 0.03$$



Variable aléatoire : Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera :

- *cas discret*, par exemple si $X \in \{0,1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face (ou client Homme/Femme, patient jeune/adulte/âgé, …).
- *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température (ou age d'un client, taille d'un patient, …).

De manière purement équivalente à la fonction de répartition $p_X(x)$, la *densité de probabilité* pourra de même représenter la loi de probabilité d'une v.a. X suivant :

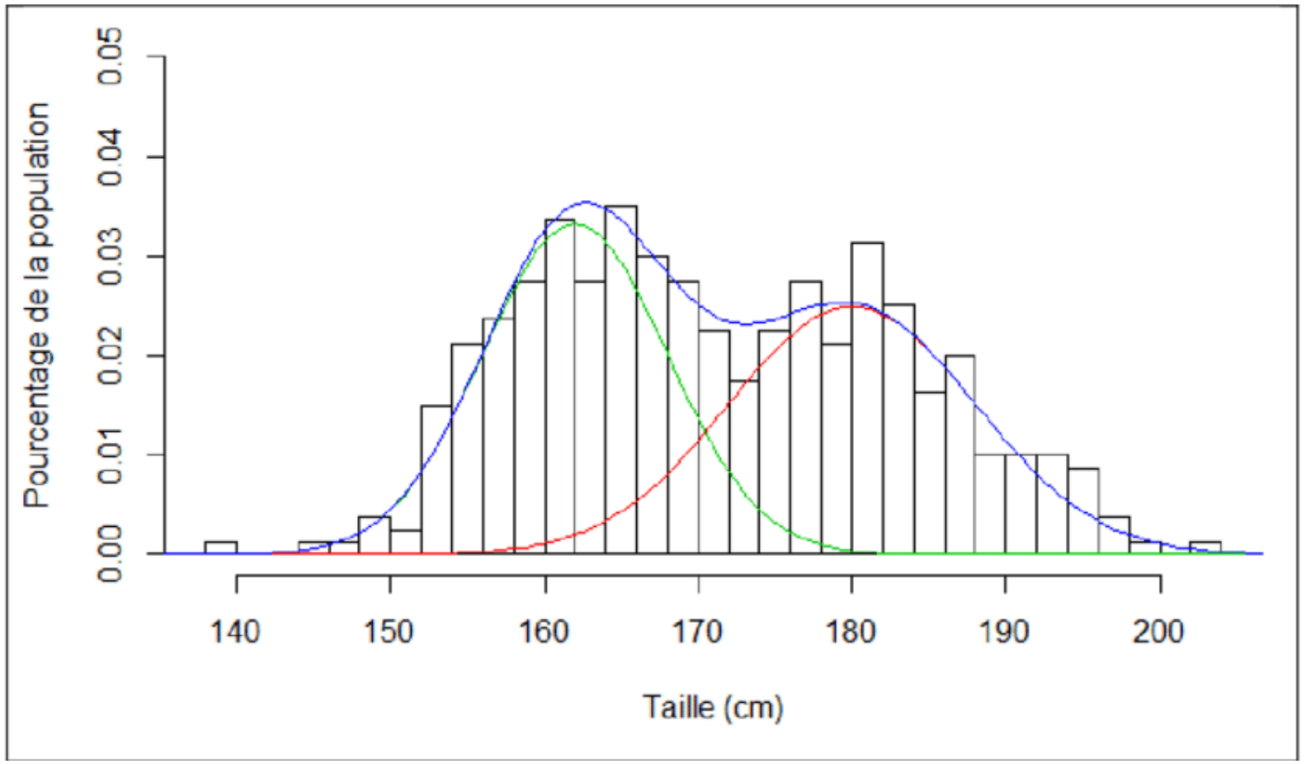
$$p_X(x) = \frac{\partial F_X}{\partial x}(x)$$

En utilisant les densités de probabilités, les chances que X tombe sur une gamme de valeurs $[x_1, x_2]$ sera alors

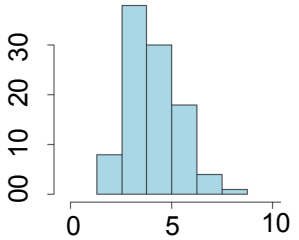
$$\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x) dx \,.$$

Variable aléatoire : Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera :

- *cas discret*, par exemple si $X \in \{0,1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face (ou client Homme/Femme, patient jeune/adulte/âgé, …).
- *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température (ou age d'un client, taille d'un patient, …).



Maintenant que les concepts mathématiques sont posés, revenons à l'exemple des « piles ou faces » avec le théorème central limite.



Afin de montrer l'importance de la loi Normale en statistique et de manipuler les concepts énoncés précédemment, il est intéressant de présenter le Théorème Central Limite (TCL).

Supposons que n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes mais suivant une même loi de probabilité soient tirés. L'espérance m et l'écart type s de leur loi est connue. Le nombre d'observations n est aussi supposé grand (typiquement $n > 30$).

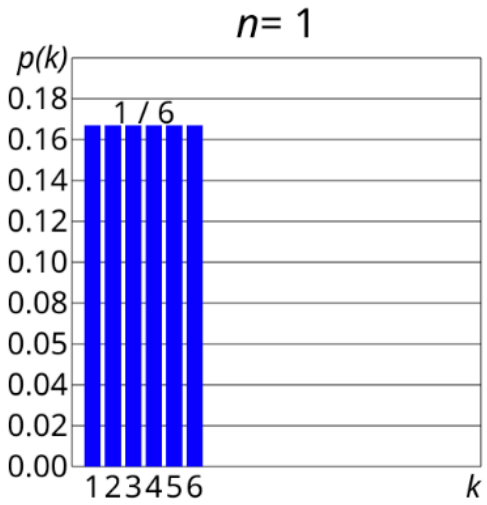
Alors, la somme des X_i peut être approchée par une loi normal de moyenne nm et d'écart type $s\sqrt{n}$:

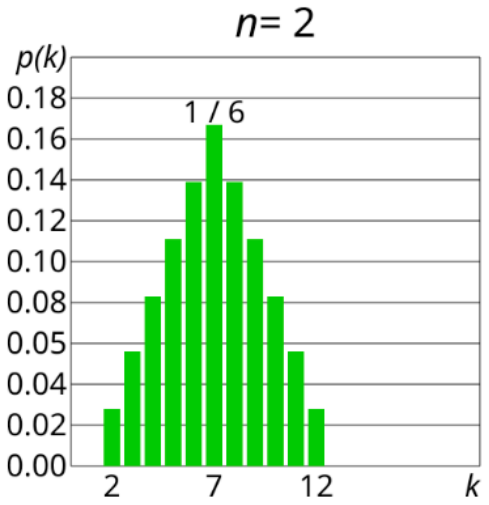
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(nm, s^2 n)\,,$$

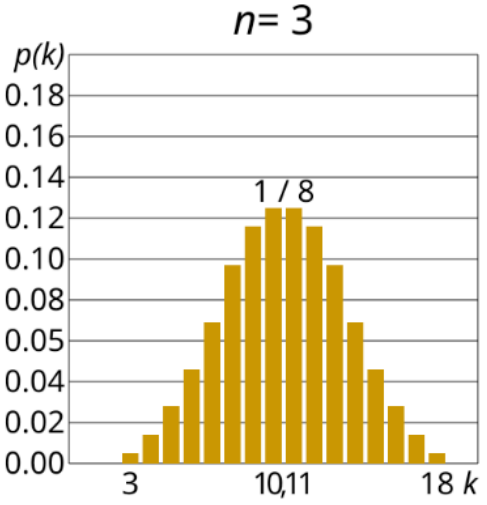
où la *densité de probabilité* de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est :

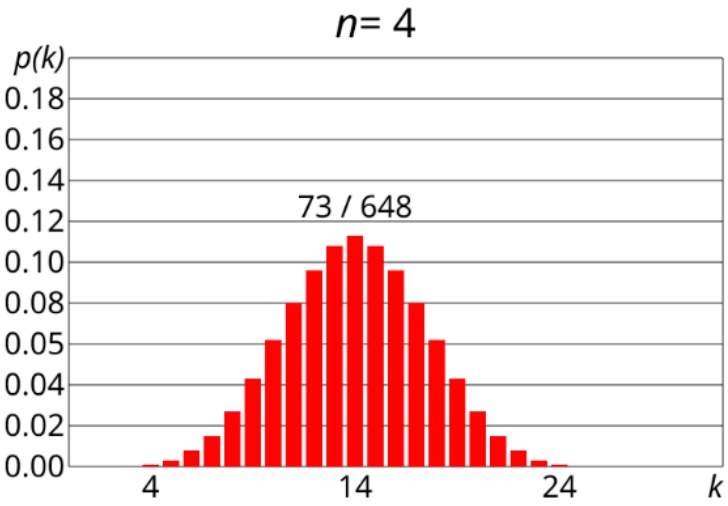
$$f_{\theta=\{\mu,\sigma\}}(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

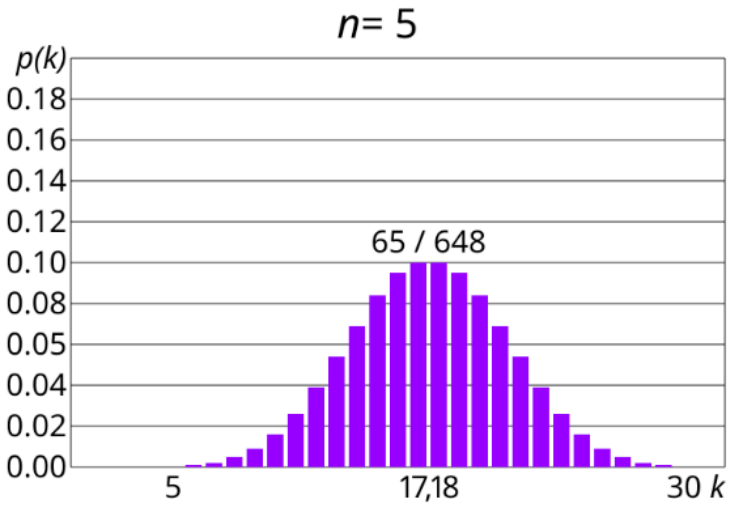
Des pièces aux dés











...

https://commons.wikimedia.org/wiki/File:Dice_sum_central_limit_theorem.svg

Au delà de la connaissance du TCL, ces expériences nous amènent plusieurs enseignements en lien avec la suite du cours :

- En assemblant plusieurs variables aléatoires, nous avons créé un modèle aléatoire qui peut inférer les valeurs de phénomènes (ici somme de plusieurs lancés de pièces ou dés)
- On peut éventuellement déterminer analytiquement différentes propriétés statistiques de ces modèles « complexes » comme leur espérance mais aussi d'autres propriétés statistiques comme leur précision/éparpillement.
- Ce type de modélisation se distingue alors de la modélisation déterministe qui ne s'intéresse qu'à l'équivalent de la moyenne ici.

Nous allons maintenant voir :

- Comment estimer les paramètres d'une loi de probabilité en fonction d'observation
- Etendre cette idée à la mise en lien de données d'entrée et de sortie observées
- Généraliser cette idée au principe de l'apprentissage automatique
- Illustrer cela à l'aide d'un modèle linéaire

Partie 2.3 : Estimation empirique des paramètres d'une loi

Nous avons vu ce qu'est une *variable aléatoire* et sa *loi*.

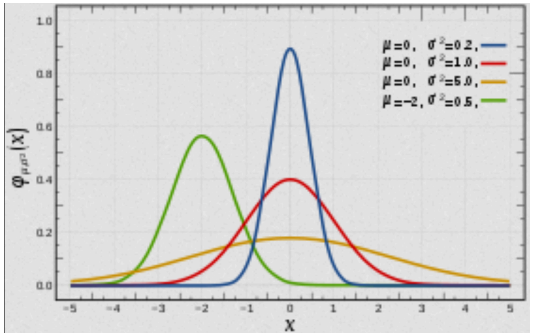
Nous avons aussi vu qu'assembler plusieurs variables permet de créer un autre objet avec ses propres propriétés en terme d'aléa (et qu'elles peuvent être étudiées).

Voyons maintenant comment *estimer les paramètres d'un modèle* représentant un phénomène *à partir d'observations* de ce phénomène.

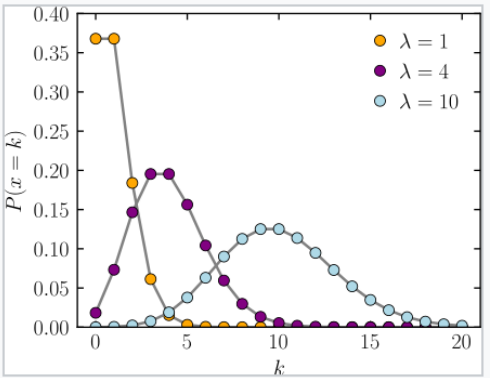
Exemple 1 : On suppose que la pièce lancée suit une loi de Bernoulli de paramètre p. Estimons p à partir de plusieurs lancés de pièces pour savoir si elle est bien équilibrée

$$\mathbb{P}(Pile) = p$$
$$\mathbb{P}(Face) = 1 - p$$

Exemple 2 : On suppose que la taille des élèves de CM2 à Toulouse suit une loi normale de moyenne m et d'écart type σ . Estimons ces paramètres à partir d'un sous-échantillon des élèves.



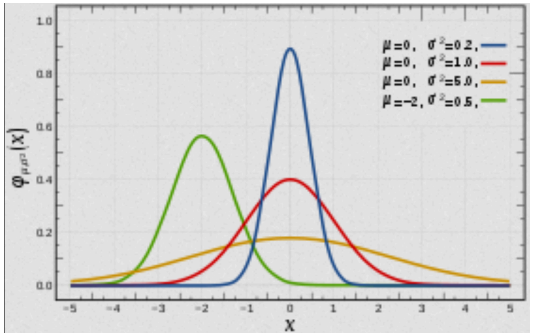
Exemple 3 : On suppose que le nombre de personnes devant nous à la caisse au magasin de journaux du quartier suit une loi de Poisson de paramètre λ. Estimons ce paramètre en observant sur quelques mois combien de personnes on avait devant nous.



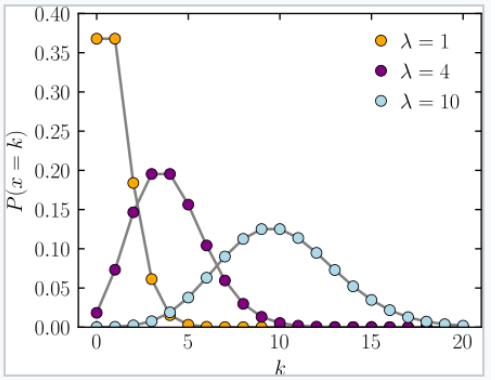
Exemple 1 : On suppose que la pièce lancée suit une loi de Bernoulli de paramètre p. Estimons p à partir de plusieurs lancés de pièces pour savoir si elle est bien équilibrée

$$\mathbb{P}(Pile) = p$$
$$\mathbb{P}(Face) = 1 - p$$

Exemple 2 : On suppose que la taille des élèves de CM2 à Toulouse suit une loi normale de moyenne m et d'écart type σ . Estimons ces paramètres à partir d'un sous-échantillon des élèves.



Exemple 3 : On suppose que le nombre de personnes devant nous à la caisse au magasin de journaux du quartier suit une loi de Poisson de paramètre λ. Estimons ce paramètre en observant sur quelques mois combien de personnes on avait devant nous.



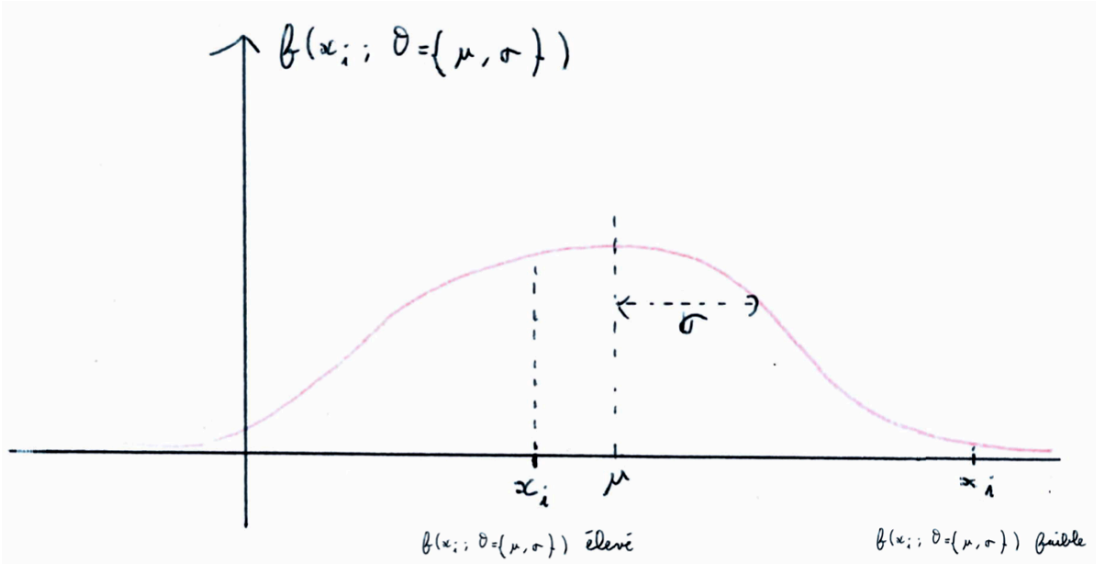
Pourquoi ces estimations ?
- Comprendre les phénomènes
- Prédire ce qui peut se passer si l'expérience est reproduite

On note :

- X une v.a. supposée suivre une loi discrète ou continue de paramètres θ.
- $x_1, \ldots, x_i, \ldots, x_n$ les observations de X.

Pour une observation x_i donnée, on modélise alors la loi de X avec la fonction $f(x_i; \theta)$ de paramètres θ (ex : moyenne et écart type d'une loi normale). Cette fonction vaut

- $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$ si X est une v.a. discrète
- $f(x_i; \theta) = f_\theta(x_i)$ si X est continue ($f_\theta(x_i)$ est la densité de probabilité de la loi)



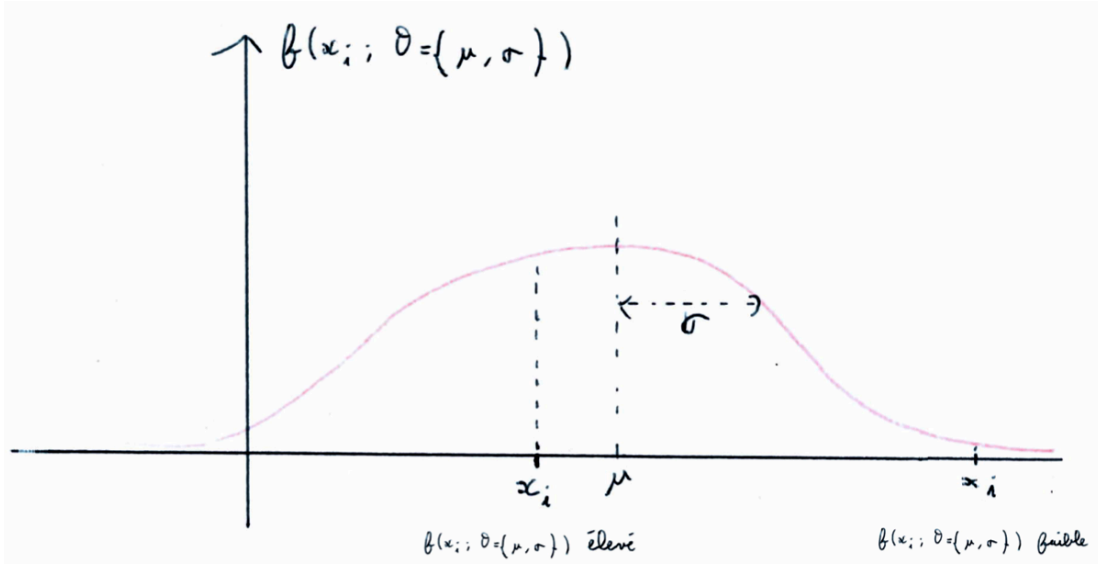
A paramètres θ fixés, $f(x_i; \theta)$ sera d'autant plus élevée que x_i a des chances d'être tirée.

On note :

- X une v.a. supposée suivre une loi discrète ou continue de paramètres θ.
- $x_1, \ldots, x_i, \ldots, x_n$ les observations de X.

Pour une observation x_i donnée, on modélise alors la loi de X avec la fonction $f(x_i; \theta)$ de paramètres θ (ex : moyenne et écart type d'une loi normale). Cette fonction vaut

- $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$ si X est une v.a. discrète
- $f(x_i; \theta) = f_\theta(x_i)$ si X est continue ($f_\theta(x_i)$ est la densité de probabilité de la loi)



A paramètres θ fixés, $f(x_i; \theta)$ sera d'autant plus élevée que x_i a des chances d'être tirée.

La vraisemblance $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ permet alors de quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.

$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.

Illustration 1 : Pile ou face

Supposons que l'on souhaite vérifier empiriquement si une pièce est équilibrée ou non.

On pose :
- $\mathbb{P}(X = 1) = f(X_i = 1; \theta = \{p\}) = p$
- $\mathbb{P}(X = 0) = f(X_i = 0; \theta = \{p\}) = 1 - p$

La vraisemblance est alors : $L(\theta = \{p\}) = \prod_{i=1}^{n} \left(1_{X_i=1} p + 1_{X_i=0} (1 - p) \right)$

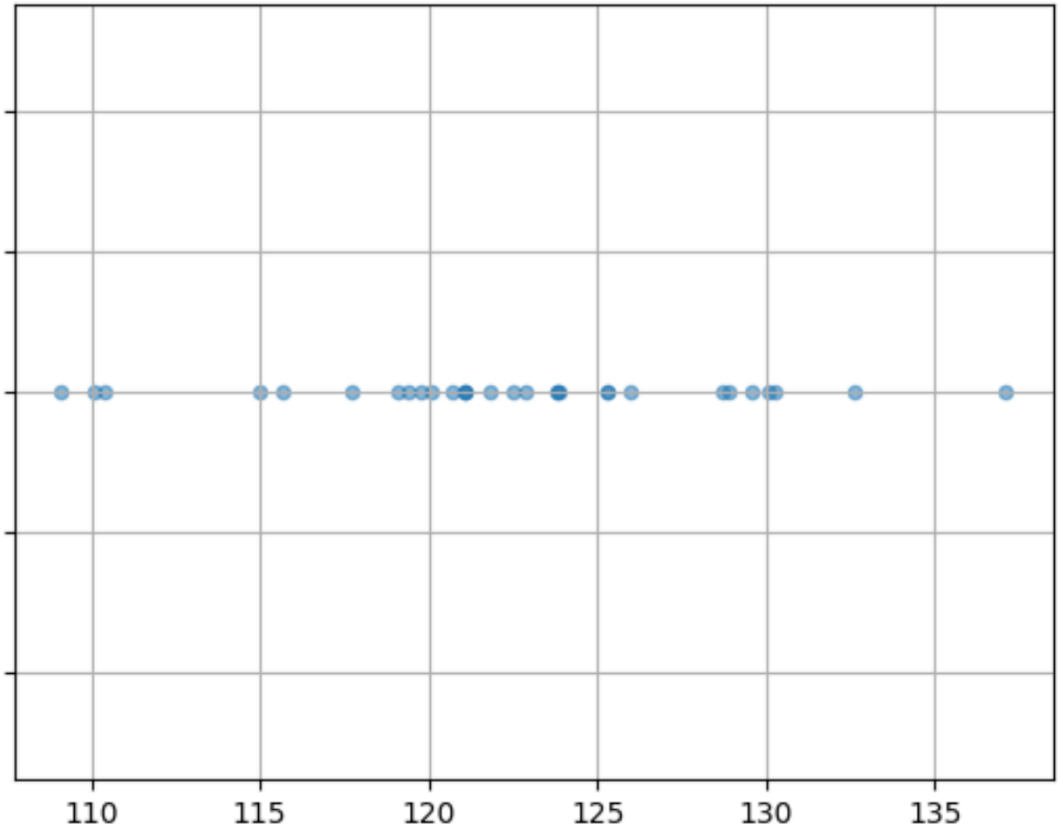
On réalise $n = 10$ observations de X et on obtient 4 'piles' et 6 'faces'.

Calculons alors la vraisemblance pour plusieurs valeurs de p :
- $L(0.2) = 0.00042$
- $L(0.5) = 0.00098$
- $L(0.8) = 0.00002$

De ces trois valeurs, $p = 0.5$ est la plus vraisemblable (mais $p = 0.4$ aurait été mieux).

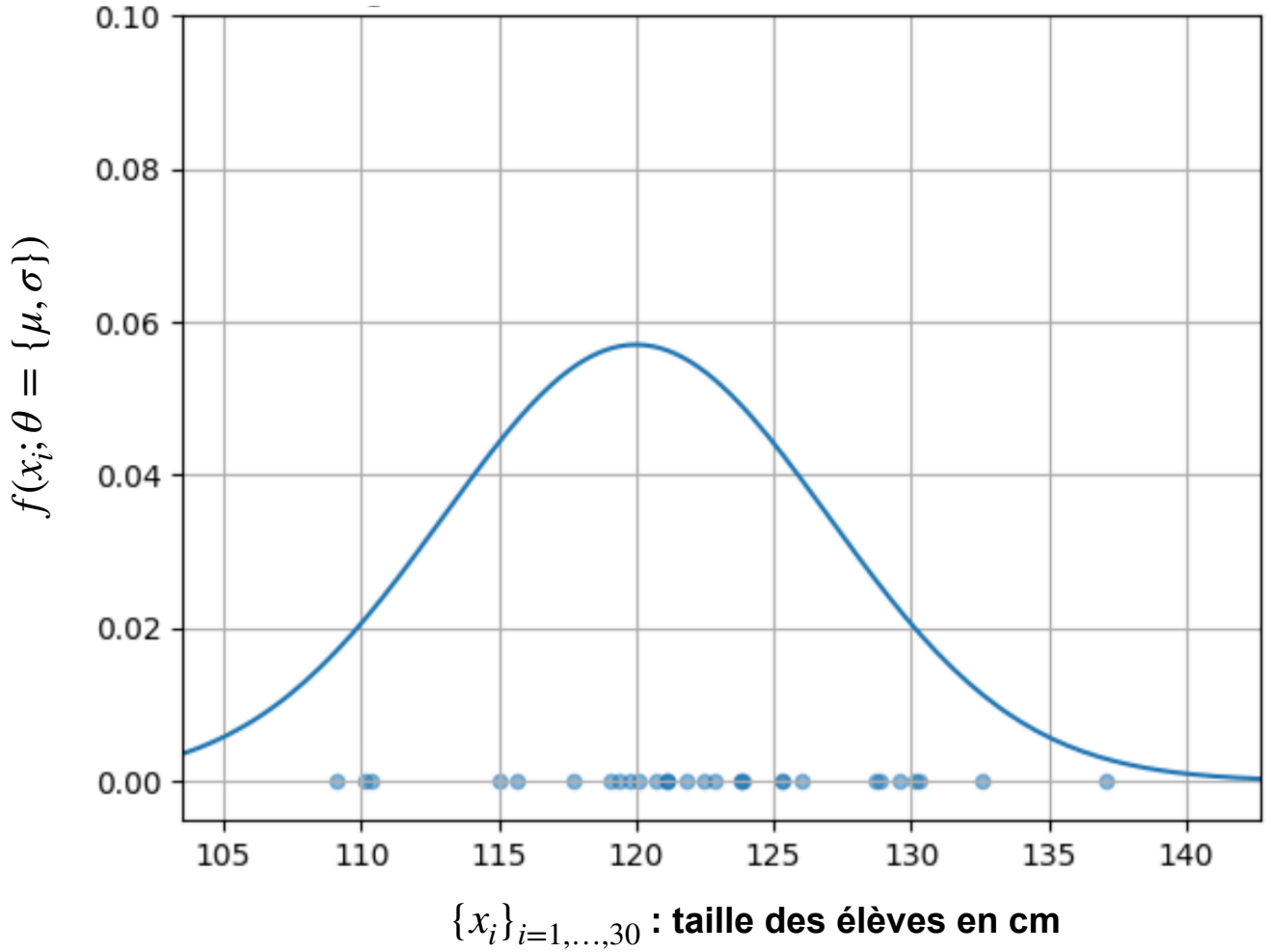
$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.

Illustration 2 : taille des élèves d'une classe de CE1



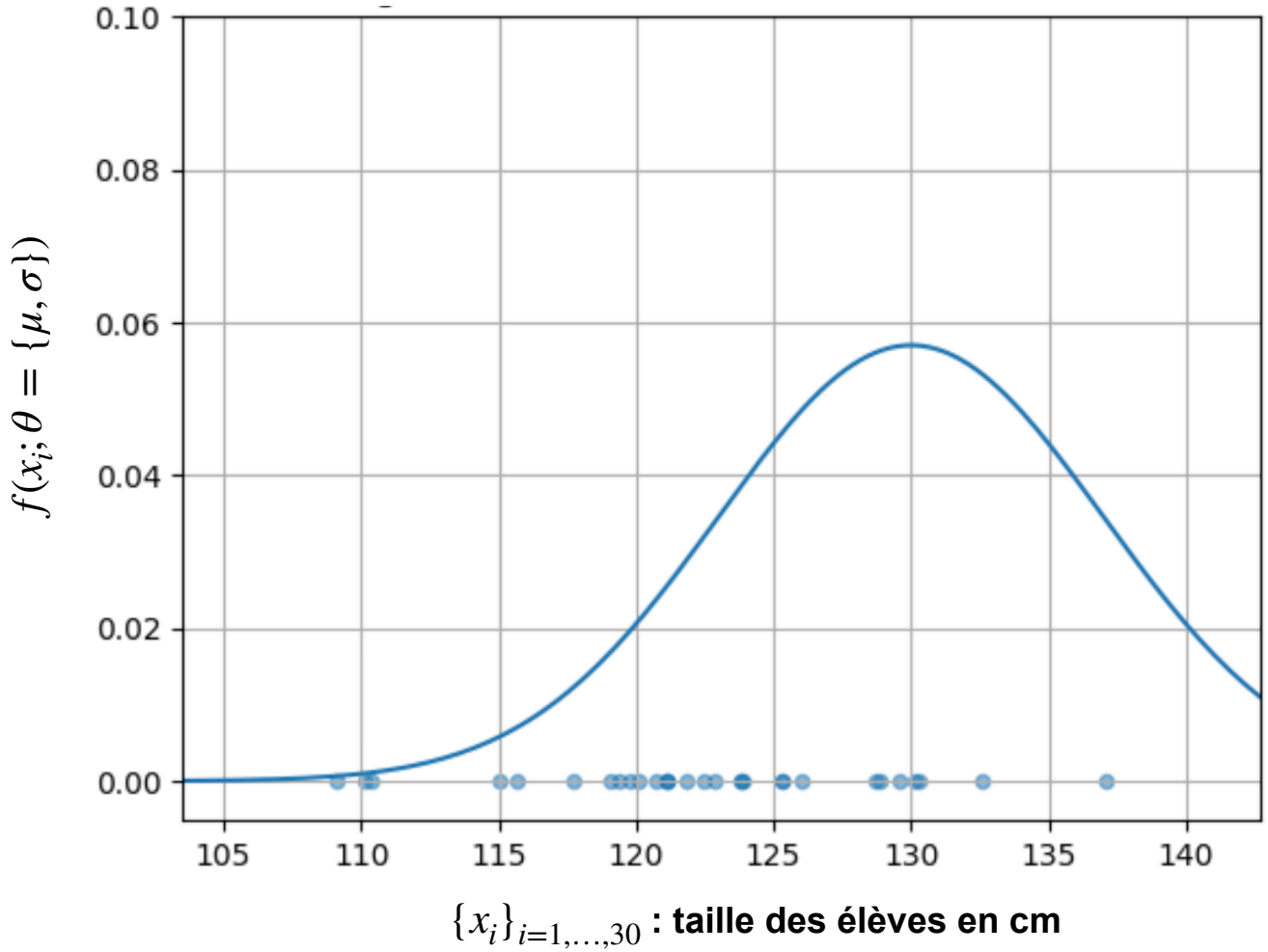
$\{x_i\}_{i=1,\ldots,30}$ **: taille des élèves en cm**

$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.



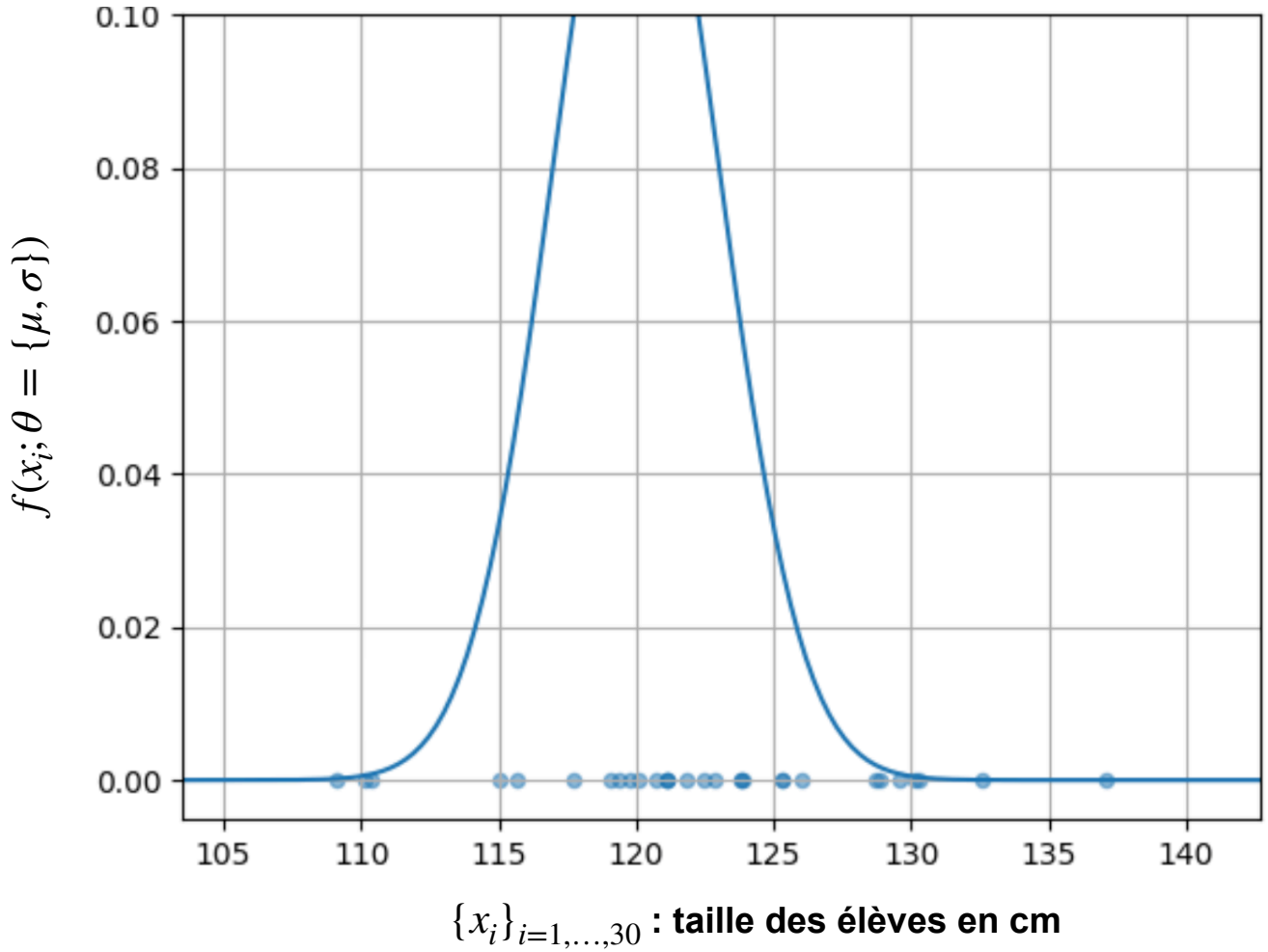
$$\log\left(L(\mu = 120\,, \sigma = 7)\right) \approx -100$$

$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.



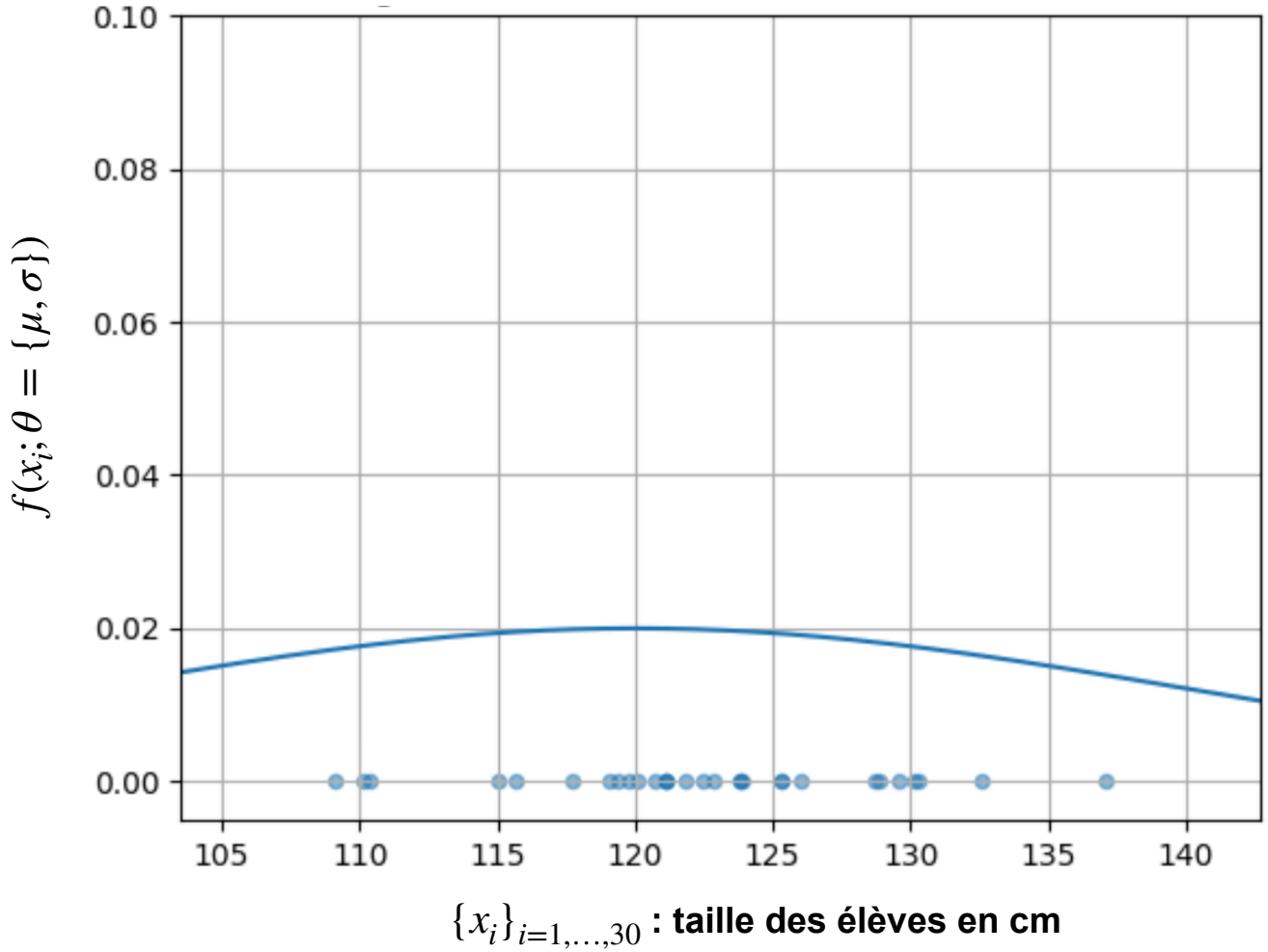
$$\log\left(L(\mu = 130\,, \sigma = 7)\right) \approx -116$$

$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.



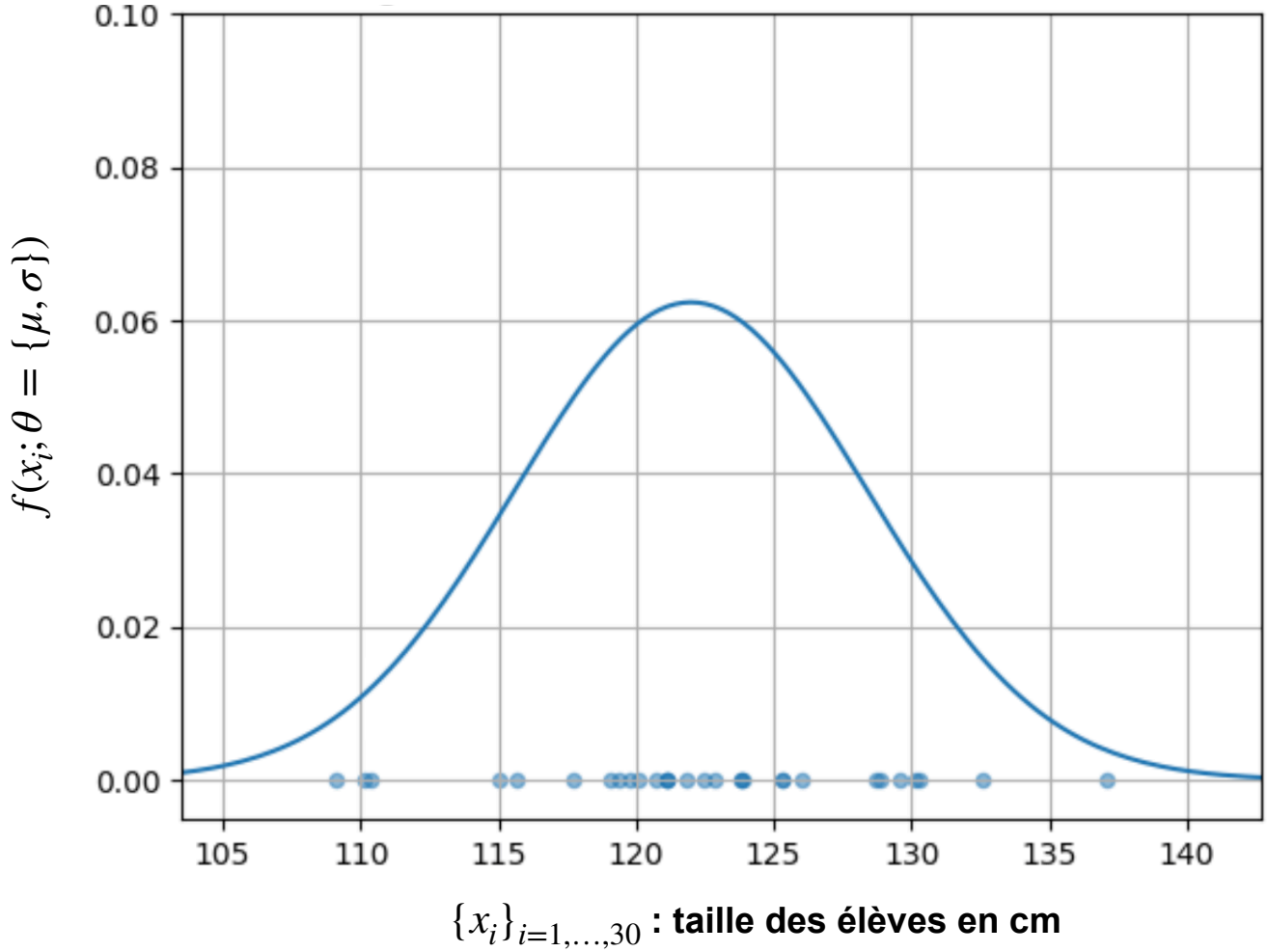
$$\log\left(L(\mu = 120\,, \sigma = 3)\right) \approx -139$$

$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.



$$\log\left(L(\mu = 120\,, \sigma = 20)\right) \approx -119$$

$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ peut quantifier l'adéquation entre des observation $x_1, \ldots, x_n$ et les paramètres θ.



Maximum de vraisemblance :

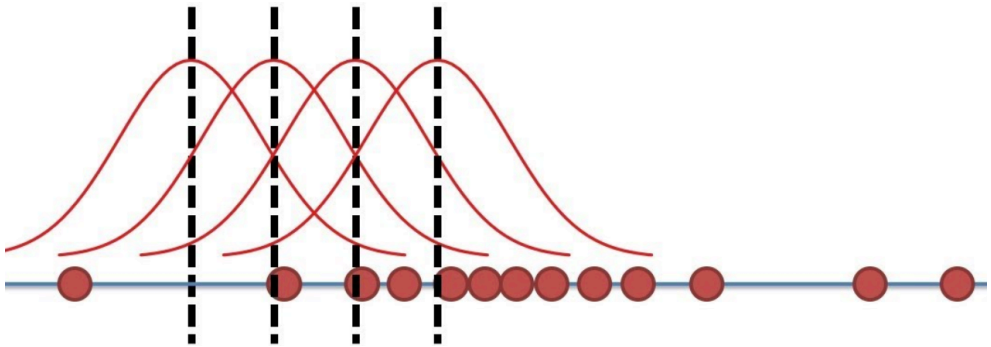
$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

$$\log\left(L(\mu = 122\,, \sigma = 6.4)\right) \approx -98$$

Pour des raisons numériques, il est aussi bien pratique de maximiser la log-vraisemblance directement, au lieu de la vraisemblance brute :

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \log\left(L(\theta)\right) \\ &= \arg\max_{\theta} \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i;\theta)\end{aligned}$$

Vu que la fonction $\log$ est strictement croissante les paramètres optimum $\hat{\theta}$ seront les mêmes avec la log-vraisemblance ou la vraisemblance.



Partie 2.4 : Apprendre une relation entre des entrées et des sorties

On connait les notes de n élèves au cours de l'année scolaire 2025-2026 ainsi que leur notes à un concours de fin d'année.

On aimerait prédire les notes au concours des élèves de la promotion 2026-2027 en fonction de leurs notes au cours de l'année.

	Maths	Info	Français	Concours
Elève 1	12	15	09	14
Elève *2*	05	09	12	07
…	…	…	…	…
Elève *i*	x_i^1	x_i^2	x_i^3	y_i
…	…	…	…	…
Elève *n*	10	12	15	11

	Maths	Info	Français	Concours
Nouvel élève	13	14	11	?

On connait les notes de n élèves au cours de l'année scolaire 2025-2026 ainsi que leur notes à un concours de fin d'année.

On aimerait prédire les notes au concours des élèves de la promotion 2026-2027 en fonction de leurs notes au cours de l'année.

	Maths	Info	Français	Concours
Elève 1	12	15	09	14
Elève *2*	05	09	12	07
…	…	…	…	…
Elève *i*	x_i^1	x_i^2	x_i^3	y_i
…	…	…	…	…
Elève *n*	10	12	15	11

	Maths	Info	Français	Concours
Nouvel élève	13	14	11	?

Posons les notations :

- Notes de l'élève $i \in \{1,\ldots,n\}$ durant l'année 2020-2021 : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right) \in \mathbb{R}^p$
- Notes de l'élève i au concours 2020-2021 : $y_i \in \mathbb{R}$
- Prédiction de y_{new} pour l'élève new en fonction des notes x_{new} : $\widehat{y_{new}} = h_\Theta(x_{new})$

Prédiction de y_{new}

Fonction $\mathbb{R}^p \to \mathbb{R}$
Paramètres Θ appris avec les $(x_i, y_i)_{i=1,\ldots,n}$

Comment apprendre le lien entre les x_i et les y_i ?

On a :

- Des données d'entrée x_i liées à des sorties y_i via un modèle $h_\theta(x_i)$
- Les x_i, y_i sont fixés et les paramètres θ sont notre inconnue
- On considère l'erreur de prédiction : $e_{i,\theta} = h_\theta(x_i) - y_i$

Comment apprendre le lien entre les x_i et les y_i ?

On a :

- Des données d'entrée x_i liées à des sorties y_i via un modèle $h_\theta(x_i)$
- Les x_i, y_i sont fixés et les paramètres θ sont notre inconnue
- On considère l'erreur de prédiction : $e_{i,\theta} = h_\theta(x_i) - y_i$

On fait l'hypothèse $e_{i,\theta} \sim \mathcal{N}(0,\sigma)$ et on maximise $L(\theta, \sigma) = \prod_{i=1}^{n} f(x_i, y_i; \theta, \sigma)$

avec $f(x_i, y_i; \theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_{i,\theta}^2}{2\sigma^2}\right)$

Comment apprendre le lien entre les x_i et les y_i ?

On a :

- Des données d'entrée x_i liées à des sorties y_i via un modèle $h_\theta(x_i)$
- Les x_i, y_i sont fixés et les paramètres θ sont notre inconnue
- On considère l'erreur de prédiction : $e_{i,\theta} = h_\theta(x_i) - y_i$

On fait l'hypothèse $e_{i,\theta} \sim \mathcal{N}(0,\sigma)$ et on maximise $L(\theta, \sigma) = \prod_{i=1}^{n} f(x_i, y_i; \theta, \sigma)$

avec $f(x_i, y_i; \theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_{i,\theta}^2}{2\sigma^2}\right)$

Plutôt que d'optimiser la vraisemblance par rapport à σ, nous allons fixer σ et optimiser :

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

Remarque : Il sera aussi possible dans certains cas d'optimiser simultanément les paramètres du modèle d'erreur (ici σ) et du modèle prédictif (ici θ)

Pourrait-on simplifier la fonction optimisée ?

Suivant hypothèse de normalité sur les erreurs $\left(h_\theta(x_i) - y_i\right) = e_{i,\theta} \sim \mathcal{N}(0,\sigma)$, le problème de maximum de vraisemblance est

$$\hat{\theta} = \arg\max_\theta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

Il est intéressant de remarquer que ce problème d'optimisation peut être ré-écrit comme :

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) \\
&= \arg\max_\theta \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e_i^2\right) \\
&= \arg\min_\theta \sum_{i=1}^{n} e_i^2 \\
&= \arg\min_\theta \sum_{i=1}^{n} (y_i - h_\Theta(x_i))^2
\end{aligned}$$

Pourrait-on simplifier la fonction optimisée ?

On a alors

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

$$= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - h_\Theta(x_i))^2$$

Cette re-formulation est très pratique car :

- Elle peut être résolue de manière analytique si h_Θ est linéaire
- Elle ouvre la porte à une stratégie d'optimisation au sens des moindre carrés :

$$\nabla_\Theta e_i^2 = 2\left(y_i - h_\Theta(x_i)\right)\nabla_\Theta h_\Theta(x_i)$$

On constatera par la suite que les techniques de minimisation de l'erreur au sens des moindres carrés sont très courantes en apprentissage automatique. En faisant écho au maximum de vraisemblance, il faudra se souvenir qu'elles reposent sur une hypothèse de normalité des erreurs du modèle prédictif optimisé.

Mise en application avec un modèle linéaire

- Utilisons un modèle très simple : La **régression linéaire** !

$$\widehat{y_i} = h_\Theta(x_i) = w_0 + \sum_{j=1}^{p} w_j x_i^j$$ dont les **paramètres** sont $\Theta = \{w_0, w_1, \ldots, w_p\}$

- Prédiction de note au concours pour un élève ayant les notes x_{new} au cours de l'année ?

Maths	Info	Français	Concours
13	14	11	?

Prenons $w_0 = 0.$, $w_1 = 0.33$, $w_2 = 0.33$ et $w_3 = 0.33$ → **Alors** $\widehat{y_{new}} = 12.54$

Mise en application avec un modèle linéaire

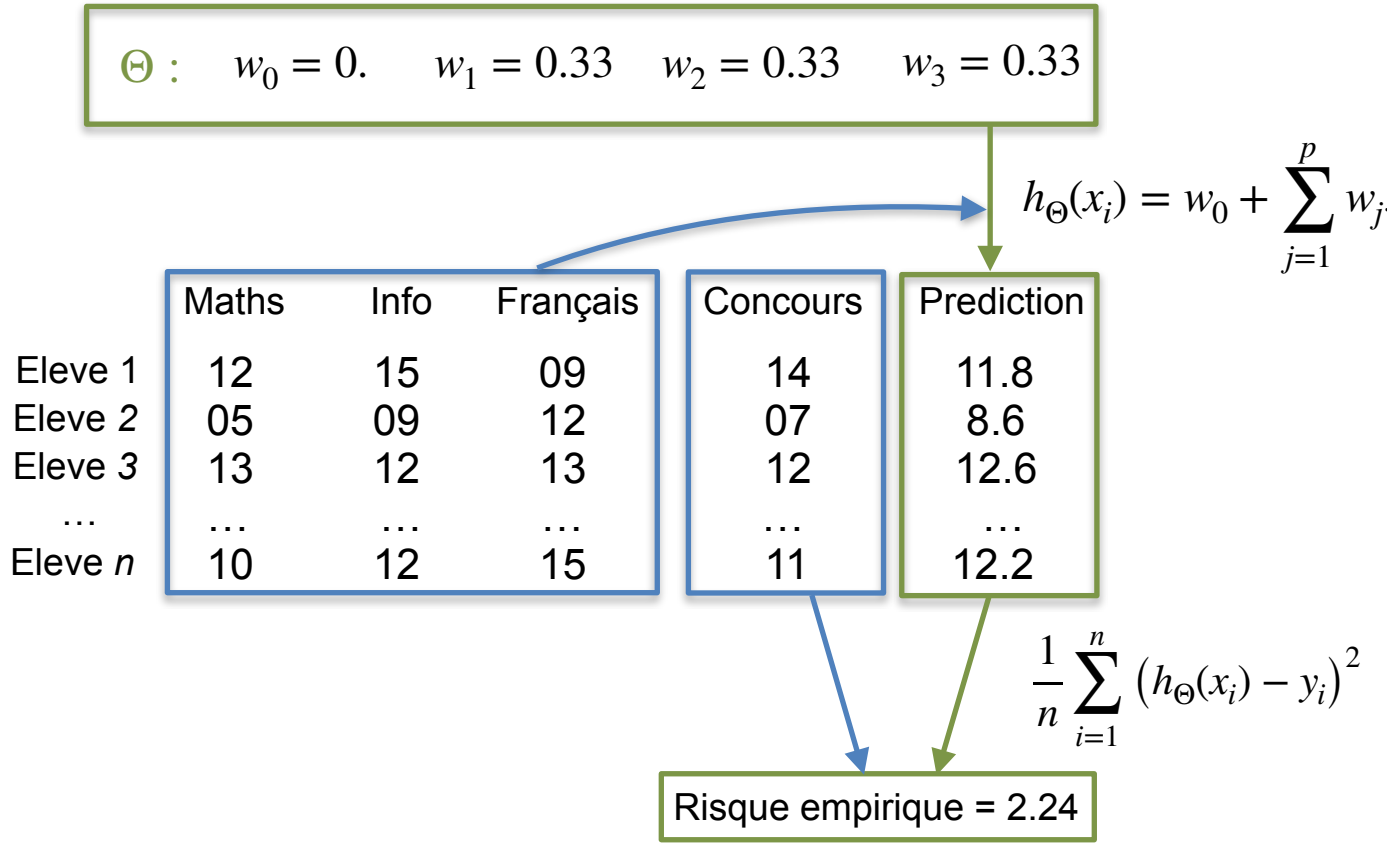
Les meilleurs $\hat{\Theta}$ minimisent la somme des erreurs de prédiction au carré des $(x_i, y_i)_{i=1,\dots,n}$:

$$\hat{\Theta} = \arg\ min_{\Theta=\{w_0,\dots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(h_\Theta(x_i) - y_i\right)^2$$

$$= \arg\ min_{\Theta=\{w_0,\dots,w_p\}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{p} w_j x_i^j - y_i \right)^2}_{\text{Risque empirique } R_\Theta((x_i, y_i)_{i=1,\dots,n})}$$

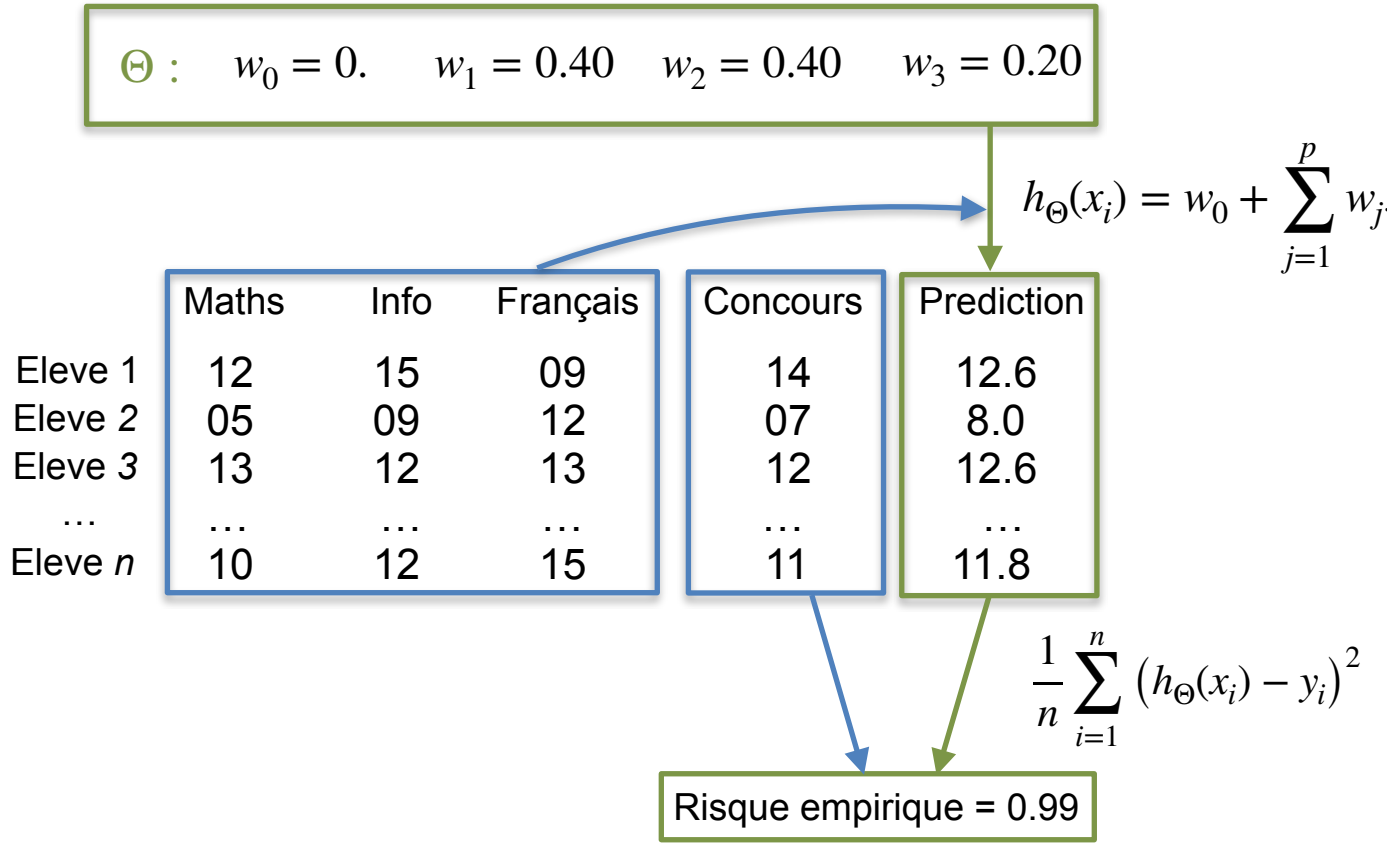
Mise en application avec un modèle linéaire

Les meilleurs $\hat{\Theta}$ minimisent la somme des erreurs de prédiction au carré des $(x_i, y_i)_{i=1,\dots,n}$:



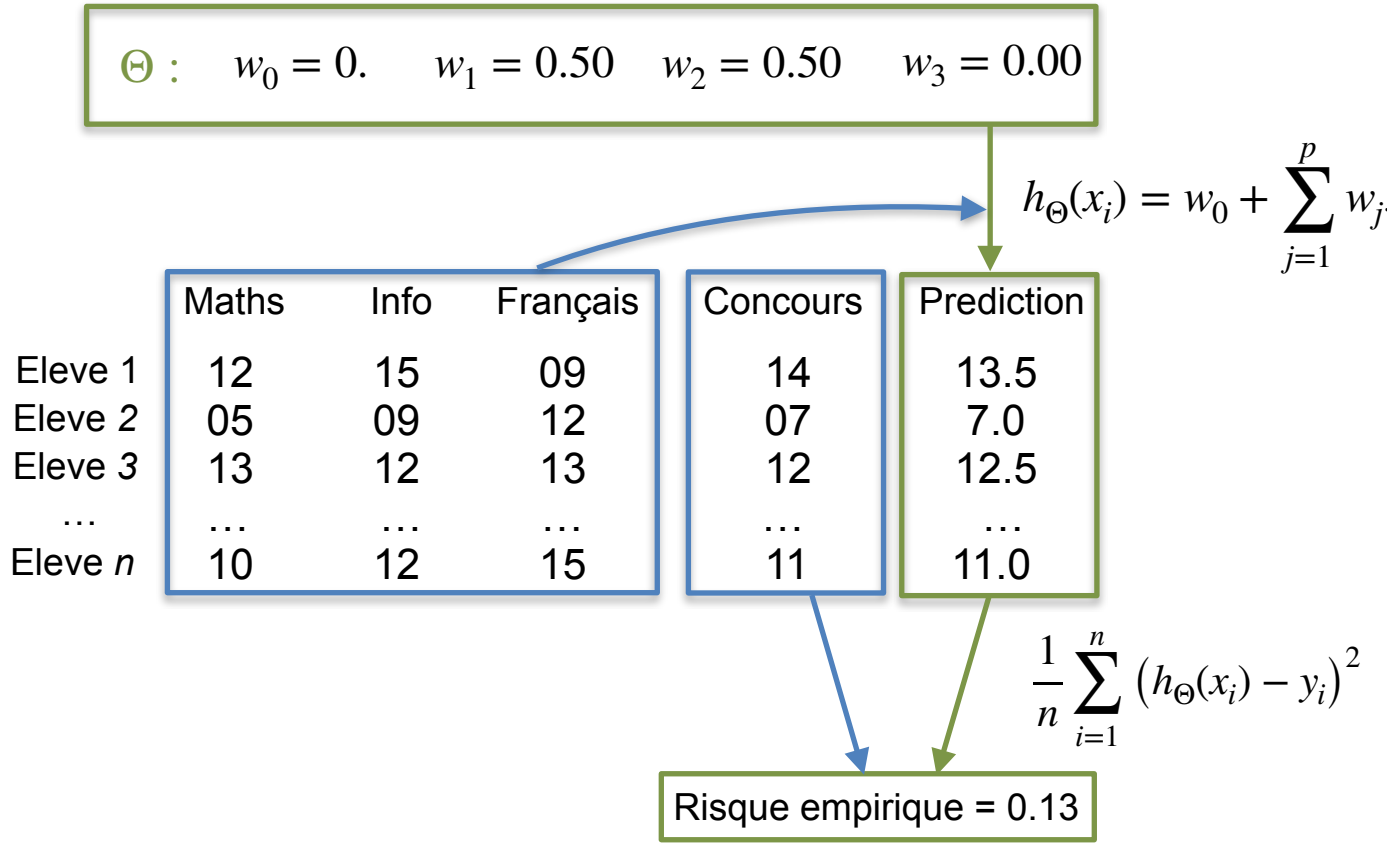
Mise en application avec un modèle linéaire

Les meilleurs $\hat{\Theta}$ minimisent la somme des erreurs de prédiction au carré des $(x_i, y_i)_{i=1,\dots,n}$:



Mise en application avec un modèle linéaire

Les meilleurs $\hat{\Theta}$ minimisent la somme des erreurs de prédiction au carré des $(x_i, y_i)_{i=1,\dots,n}$:



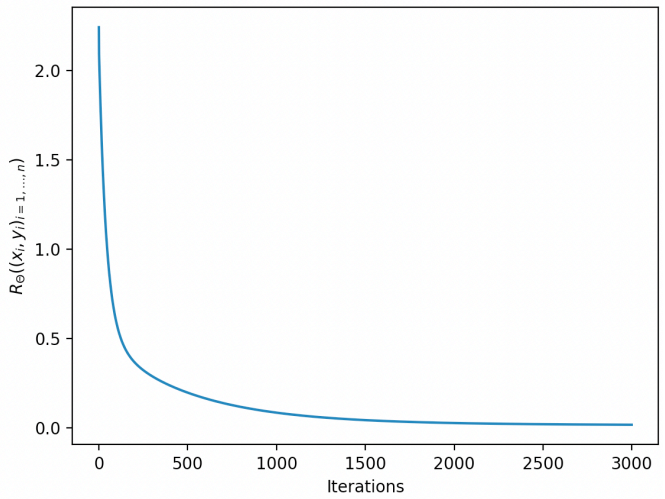
Mise en application avec un modèle linéaire

Les meilleurs $\hat{\Theta}$ minimisent la somme des erreurs de prédiction au carré des $(x_i, y_i)_{i=1,\dots,n}$:

Résolution automatique

- Formulation analytique possible ici pour estimer le résultat optimal (zéros du gradient).
- Sinon utilisation de la descente de gradient (G.D.)

$$\nabla_\Theta R_\Theta(\dots) = \left(\frac{\partial R_\Theta(\dots)}{\partial w_0}, \frac{\partial R_\Theta(\dots)}{\partial w_1}, \dots, \frac{\partial R_\Theta(\dots)}{\partial w_p} \right)^\intercal = \frac{1}{n} \sum_{i=1}^{n} 2 \left(h_\Theta(x_i) - y_i \right) \left(1, x_i^1, \dots, x_i^p \right)^\intercal$$



On trouve :

- Analytiquement : $(w_0, w_1, w_2, w_3)^\intercal \approx (-0.00, 0.5, 0.5, 0.0)^\intercal$
- Par G.D. : $(w_0, w_1, w_2, w_3)^\intercal \approx (-0.00, 0.48, 0.53, -0.01)^\intercal$

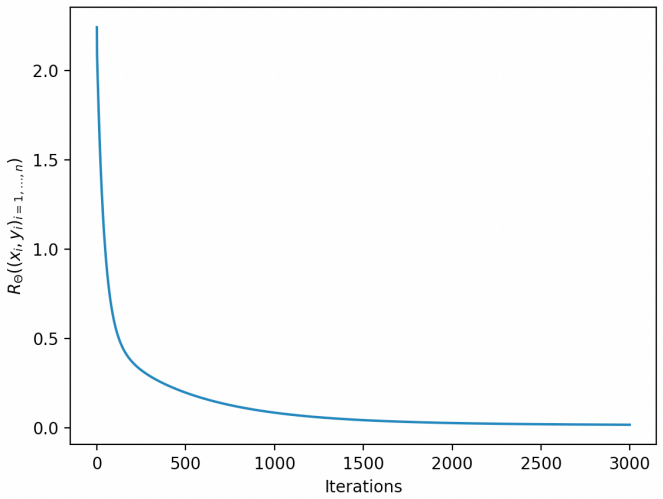
Mise en application avec un modèle linéaire

Les meilleurs $\hat{\Theta}$ minimisent la somme des erreurs de prédiction au carré des $(x_i, y_i)_{i=1,\dots,n}$:

Résolution automatique

- Formulation analytique possible ici pour estimer le résultat optimal (zéros du gradient).
- Sinon utilisation de la descente de gradient (G.D.)

$$\nabla_\Theta R_\Theta(\dots) = \left(\frac{\partial R_\Theta(\dots)}{\partial w_0}, \frac{\partial R_\Theta(\dots)}{\partial w_1}, \dots, \frac{\partial R_\Theta(\dots)}{\partial w_p} \right)^\intercal = \frac{1}{n} \sum_{i=1}^{n} 2 \left(h_\Theta(x_i) - y_i \right) \left(1, x_i^1, \dots, x_i^p \right)^\intercal$$



On trouve :

- Analytiquement : $(w_0, w_1, w_2, w_3)^\intercal \approx (-0.00, 0.5, 0.5, 0.0)^\intercal$
- Par G.D. : $(w_0, w_1, w_2, w_3)^\intercal \approx (-0.00, 0.48, 0.53, -0.01)^\intercal$

Et nous pouvons faire des prédictions sur de nouvelles observations :

Maths	Info	Français	Concours
13	14	11	**13.5**

Partie 2.5 : Apprendre et bien généraliser

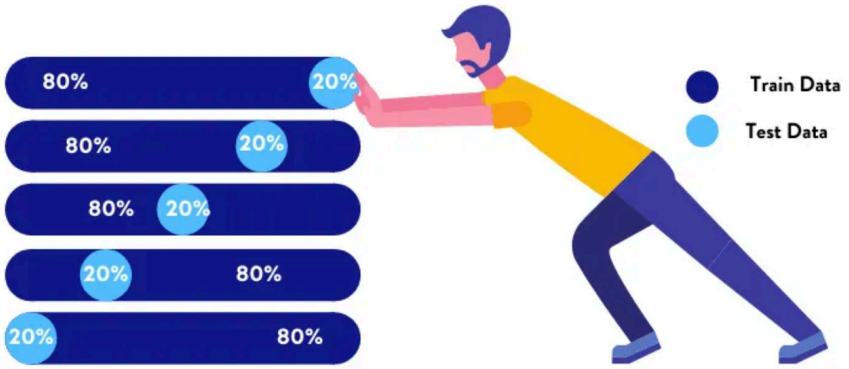
Apprentissage automatique :

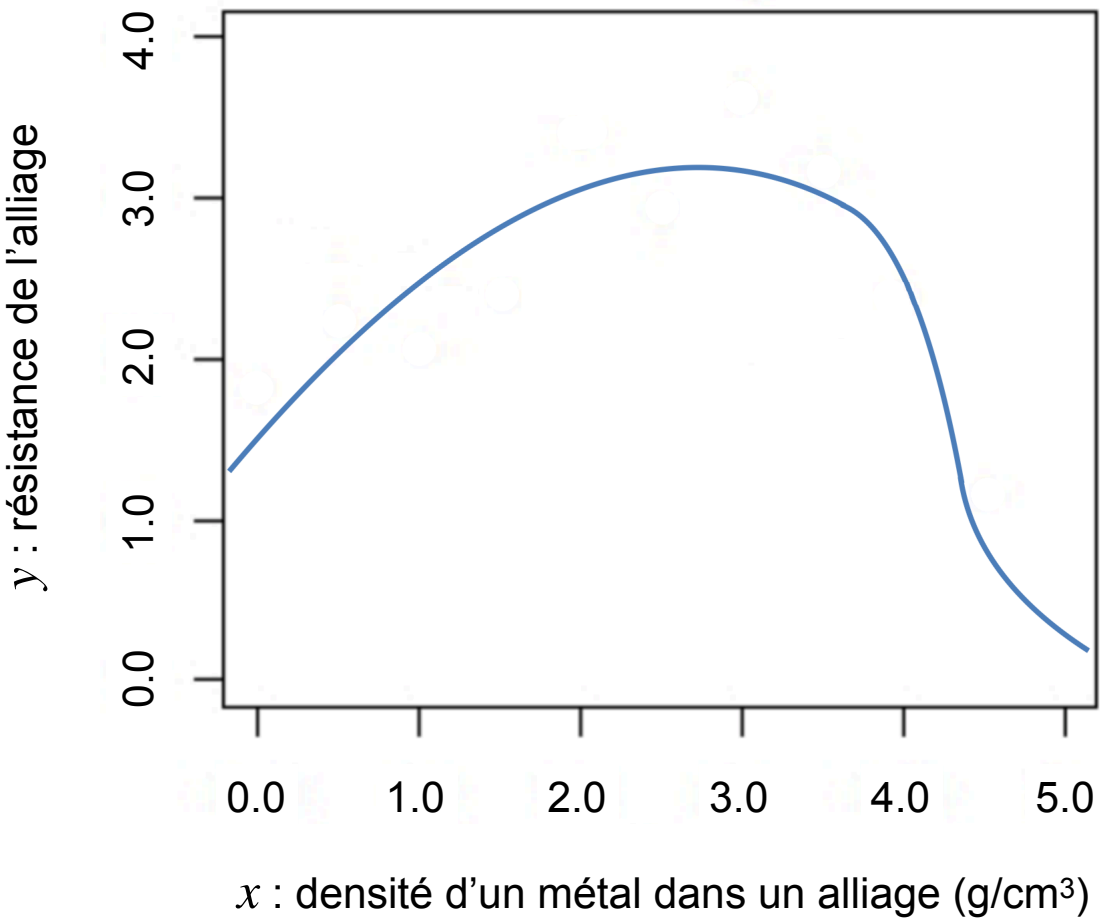
- Le premier principe clé de l'apprentissage automatique est d'optimiser les règles de décisions à partir d'exemples dit d'apprentissage (technique et demandeur en ressources).
- Les prédictions sont alors appliquées à de nouvelles données (facile et rapide).

Question :

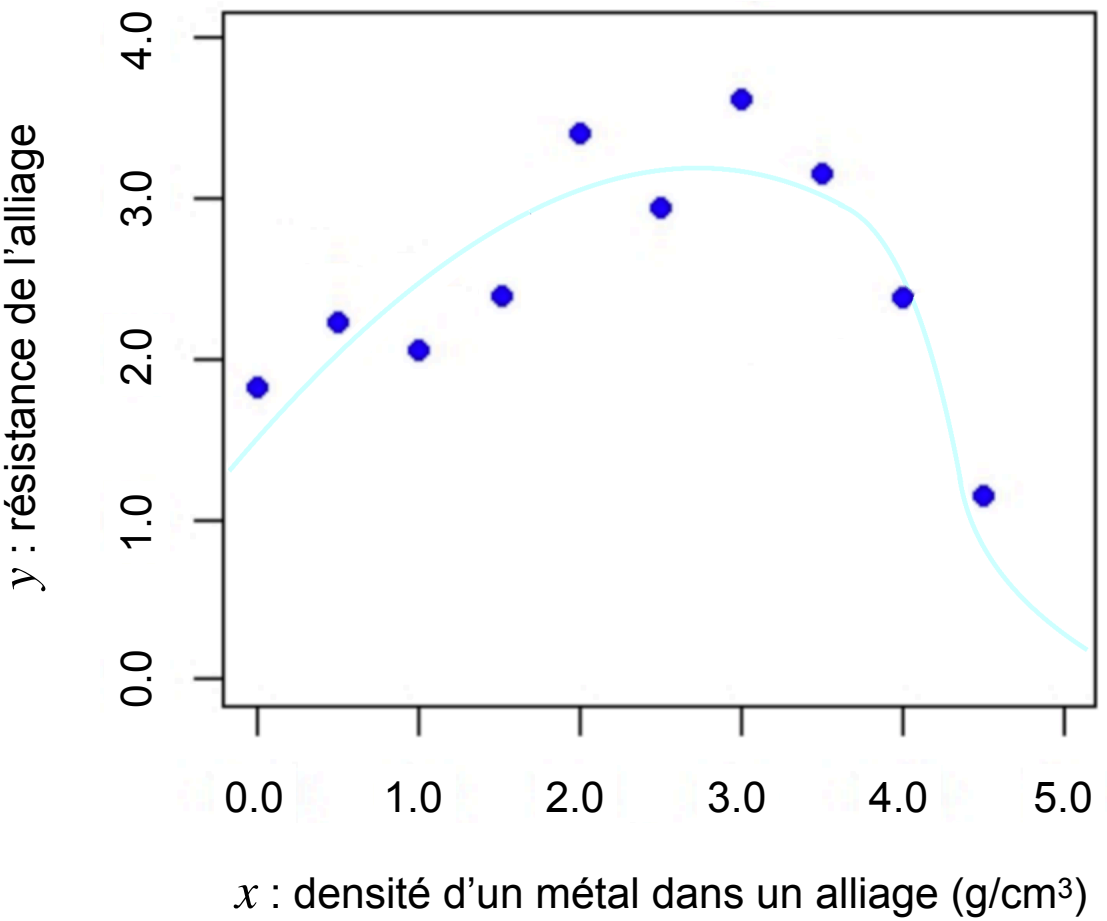
Comment s'assurer que les prédictions se généralisent bien ?

→ Deuxième principe clé de l'apprentissage automatique : la validation croisée





Relation inconnue entre une variable x et une variable y



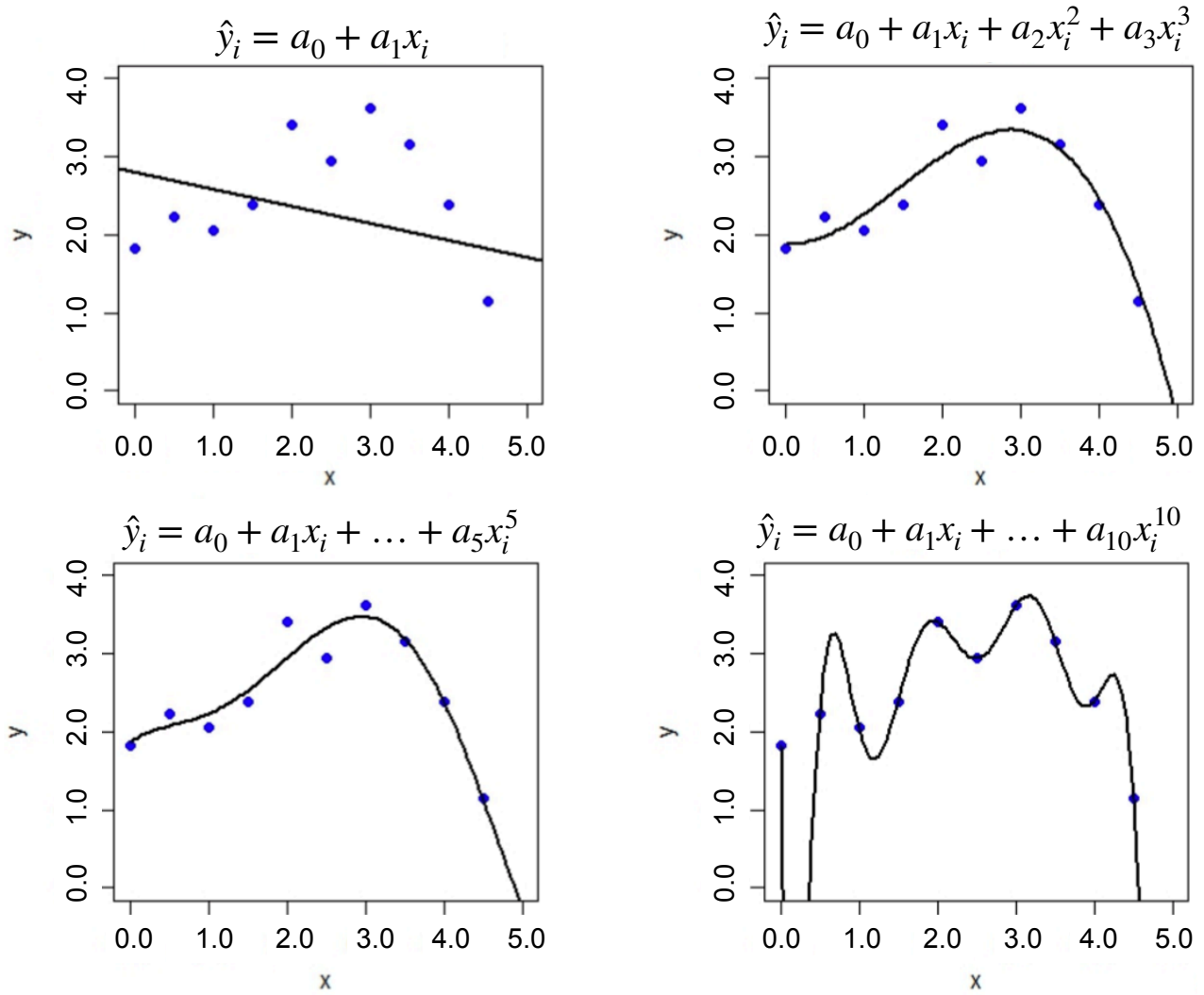
Motivation :

- On peut contrôler x
- y est difficile à obtenir

Approche M.L. :

- On observe des $(x_i, y_i)_{i=1,\dots,n}$
- On apprend un modèle $f_\theta(x)$ tel que ses prédictions $\hat{y}_i = f_\theta(x_i)$ sont proches des y_i

Compromis biais-variance



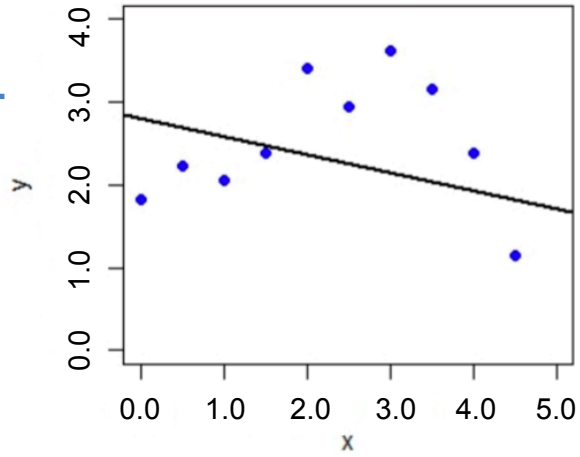
Modèles polynomiaux optimaux, avec $\{2, 4, 6, 11\}$ degrés de liberté

→ Fort impact du choix du modèle !

Compromis biais-variance

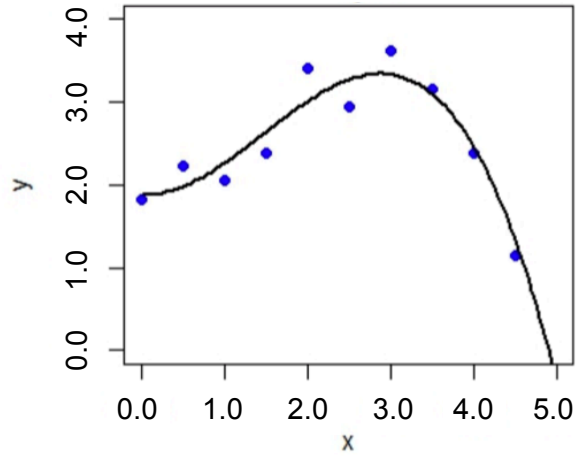
$\hat{y}_i = a_0 + a_1 x_i$

$R^2 \approx 0.11$ ←



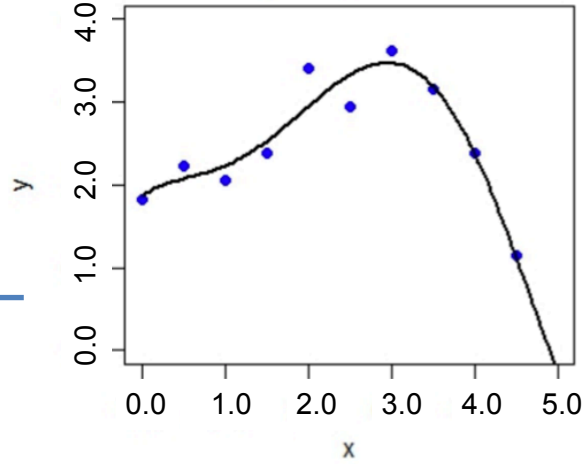
$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3$

→ $R^2 \approx 0.95$

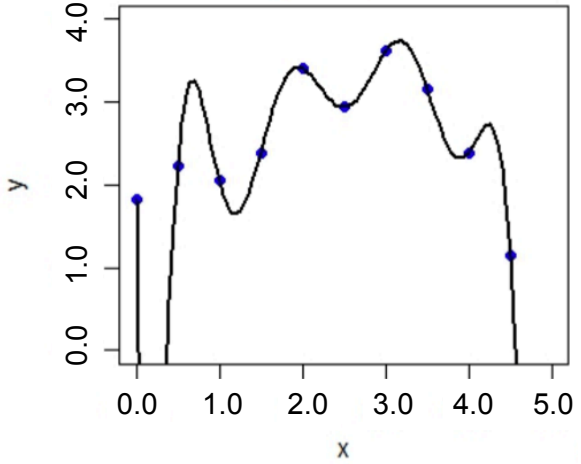


$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_5 x_i^5$

$R^2 \approx 0.96$ ←



$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_{10} x_i^{10}$



→ $R^2 \approx 1.0$

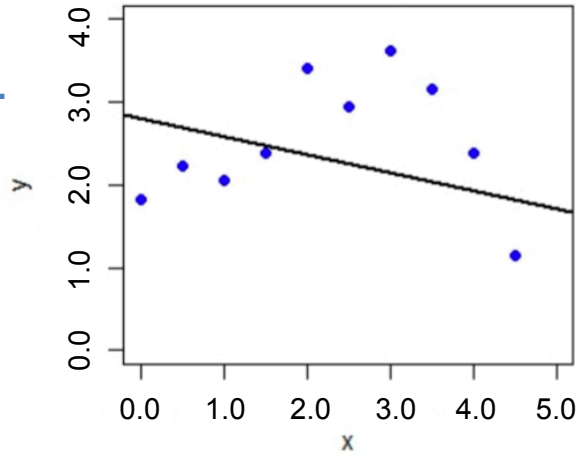
Utilisation du score $R^2 = 1 - \dfrac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}$ où $\bar{y}$ est le y_i moyen :

→ indique à quel point la variabilité des données capturée par le modèle

Compromis biais-variance

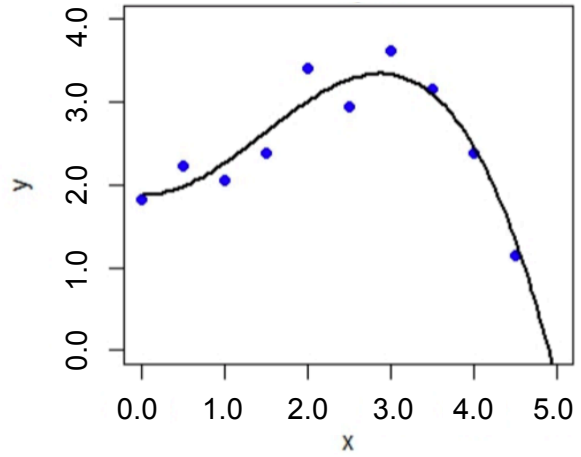
$\hat{y}_i = a_0 + a_1 x_i$

$R^2 \approx 0.11$ ←



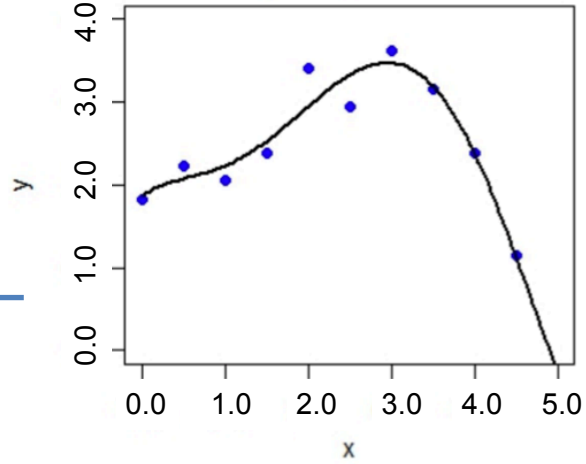
$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3$

→ $R^2 \approx 0.95$

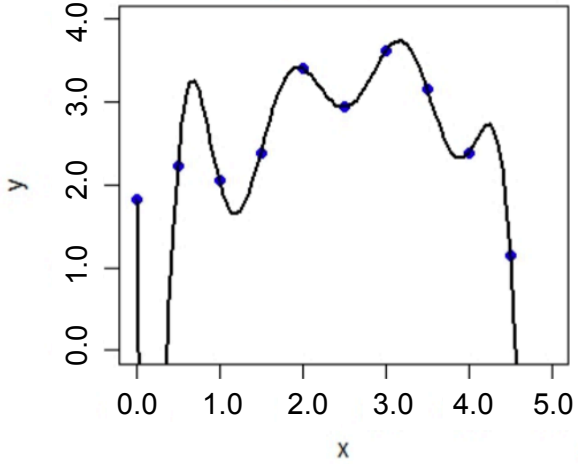


$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_5 x_i^5$

$R^2 \approx 0.96$ ←



$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_{10} x_i^{10}$



→ $R^2 \approx 1.0$

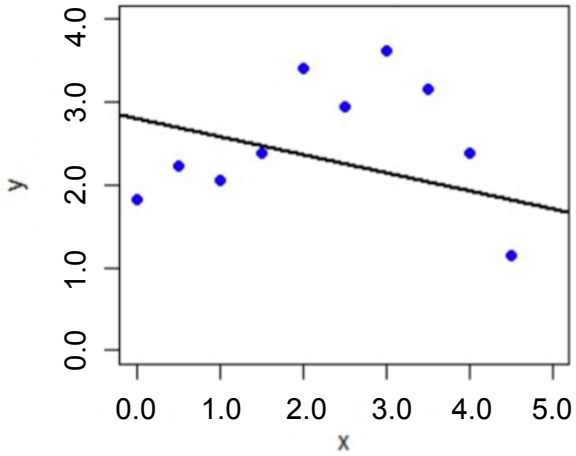
Utilisation du score $R^2 = 1 - \dfrac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}$ où $\bar{y}$ est le y_i moyen :

→ indique à quel point la variabilité des données capturée par le modèle

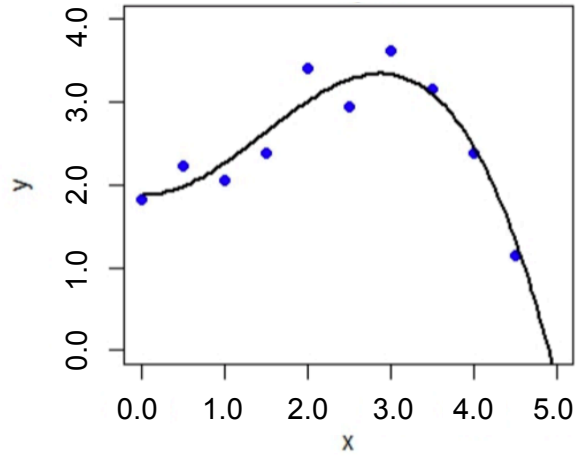
Ne tient pas compte des capacités du modèle à généraliser hors des données d'apprentissage.

Compromis biais-variance

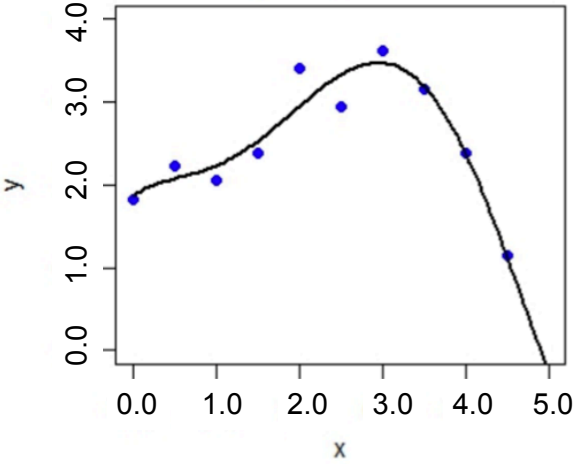
$\hat{y}_i = a_0 + a_1 x_i$



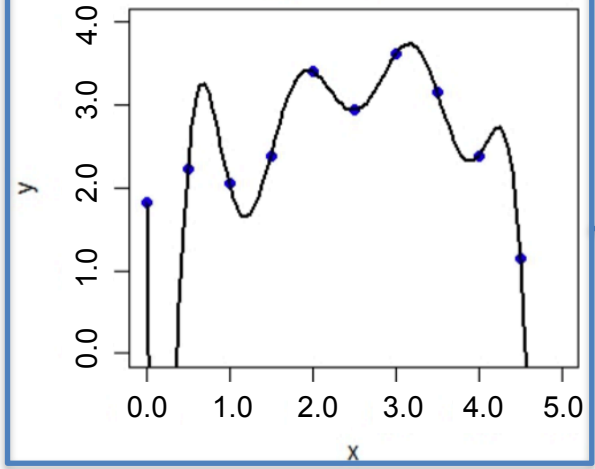
$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3$



$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_5 x_i^5$



$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_{10} x_i^{10}$



Faible biais : Les prédictions aux points d'apprentissage sont parfaites

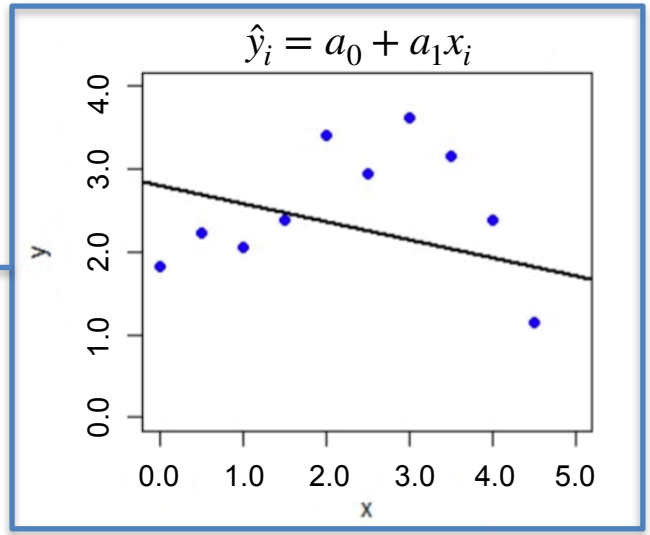
Forte variance : Mauvaise généralisation des prédictions en dehors des points d'apprentissage

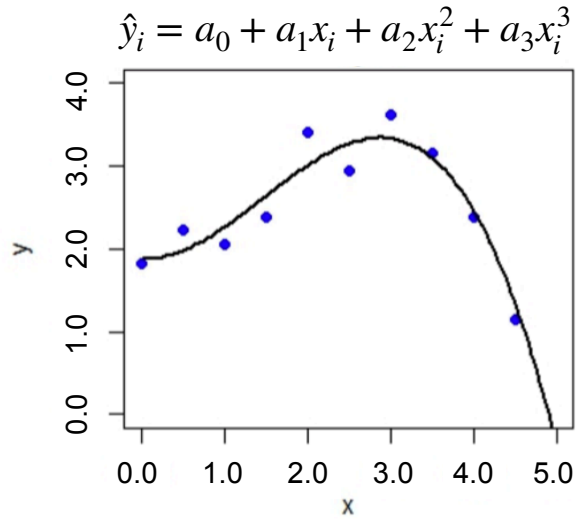
Sur-apprentissage

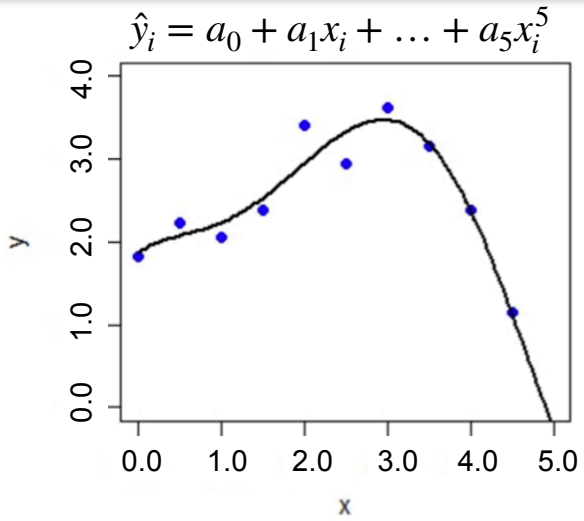
Compromis biais-variance

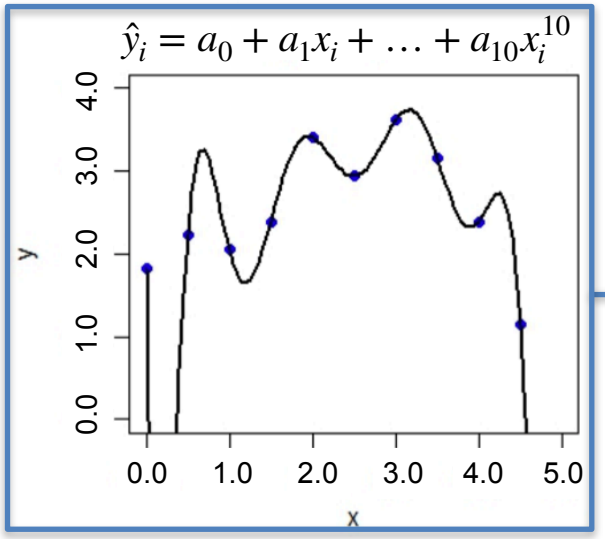
Fort biais :
Les prédictions aux points d'apprentissage sont mauvaises

Faible variance :
Bonne généralisation des prédictions en dehors des points d'apprentissage









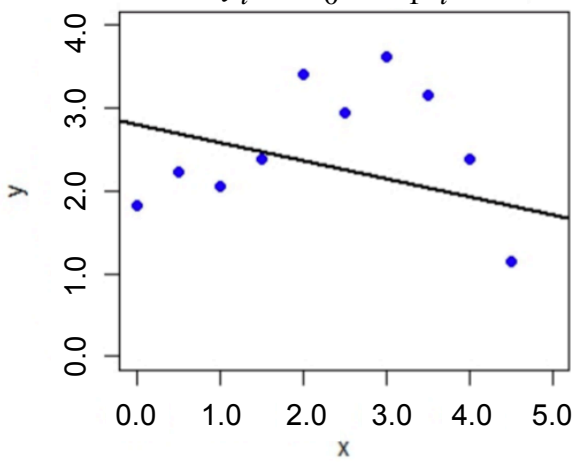
Faible biais :
Les prédictions aux points d'apprentissage sont parfaites

Forte variance :
Mauvaise généralisation des prédictions en dehors des points d'apprentissage

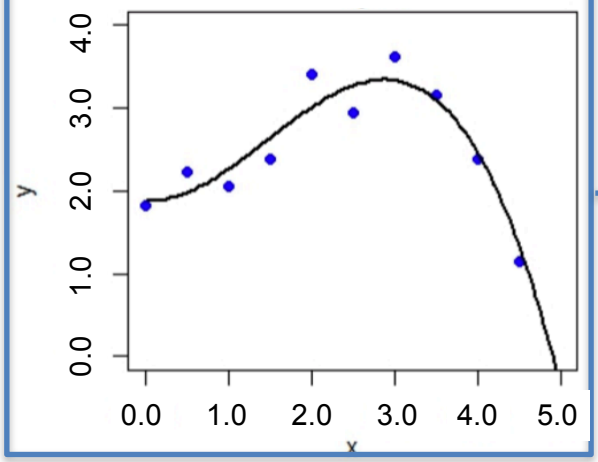
Sous-apprentissage

Compromis biais-variance

$\hat{y}_i = a_0 + a_1 x_i$

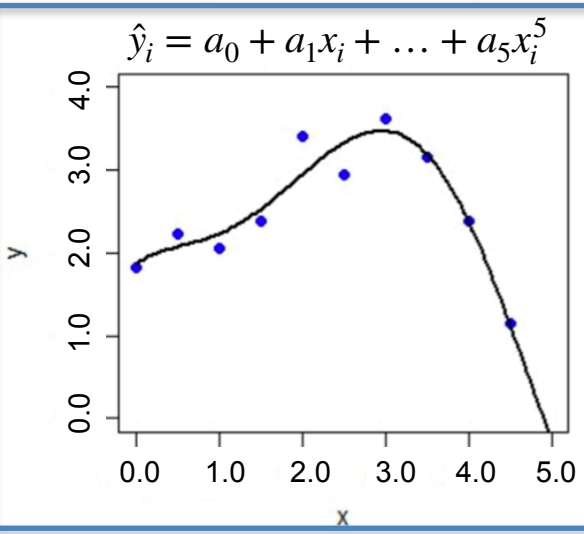


$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3$



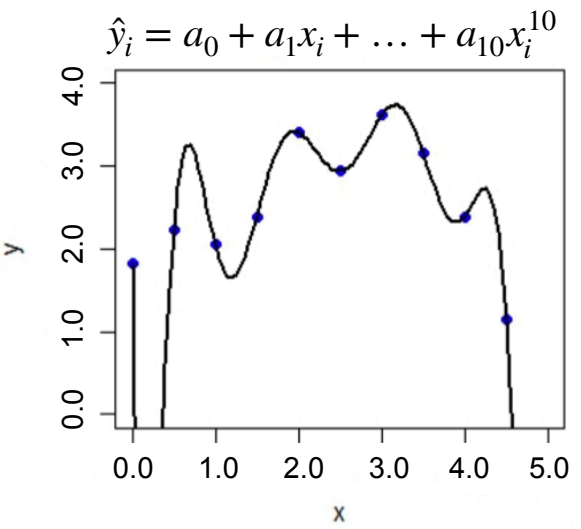
Bon compromis biais-variance

$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_5 x_i^5$

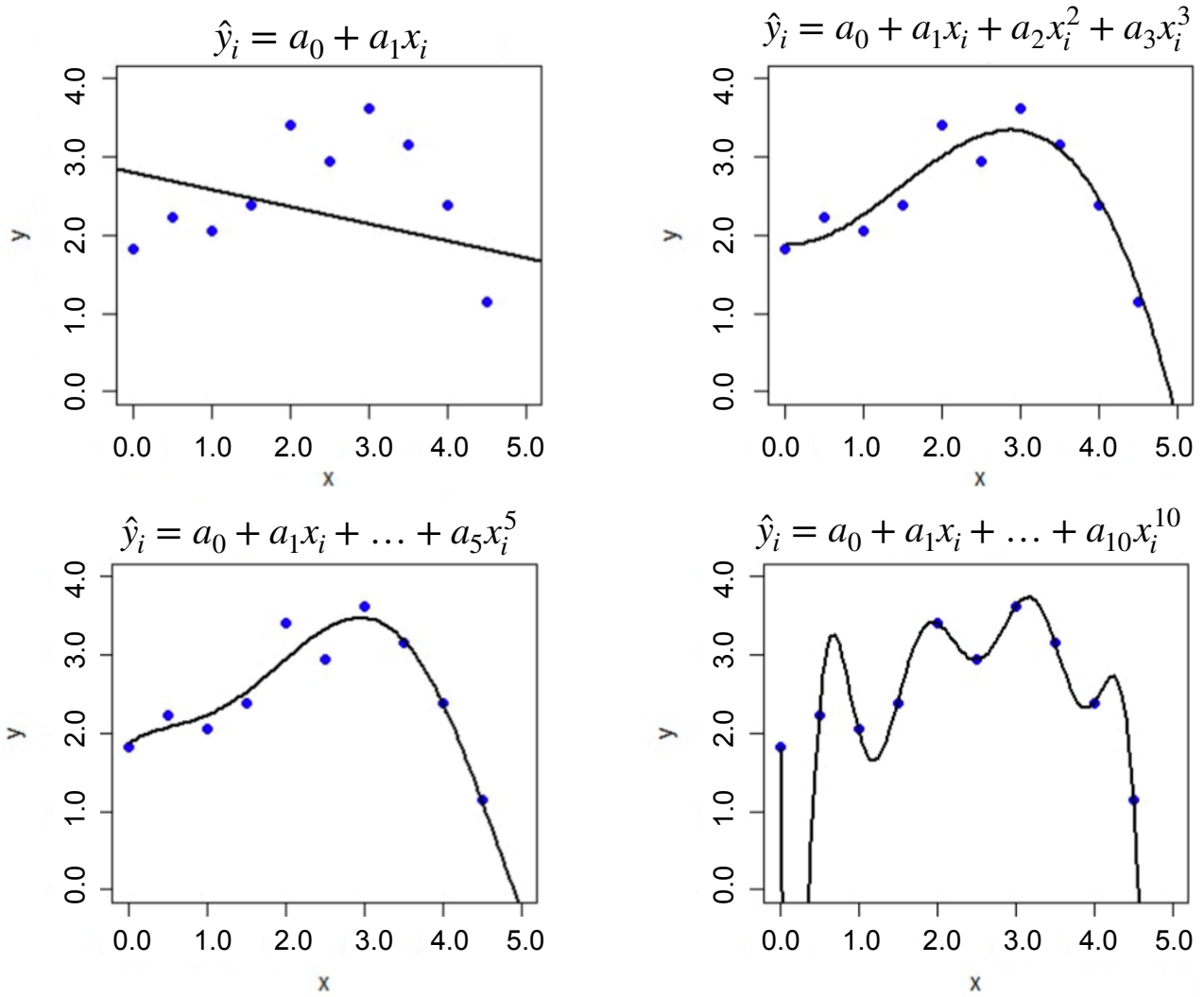


Bon compromis biais-variance

$\hat{y}_i = a_0 + a_1 x_i + \ldots + a_{10} x_i^{10}$



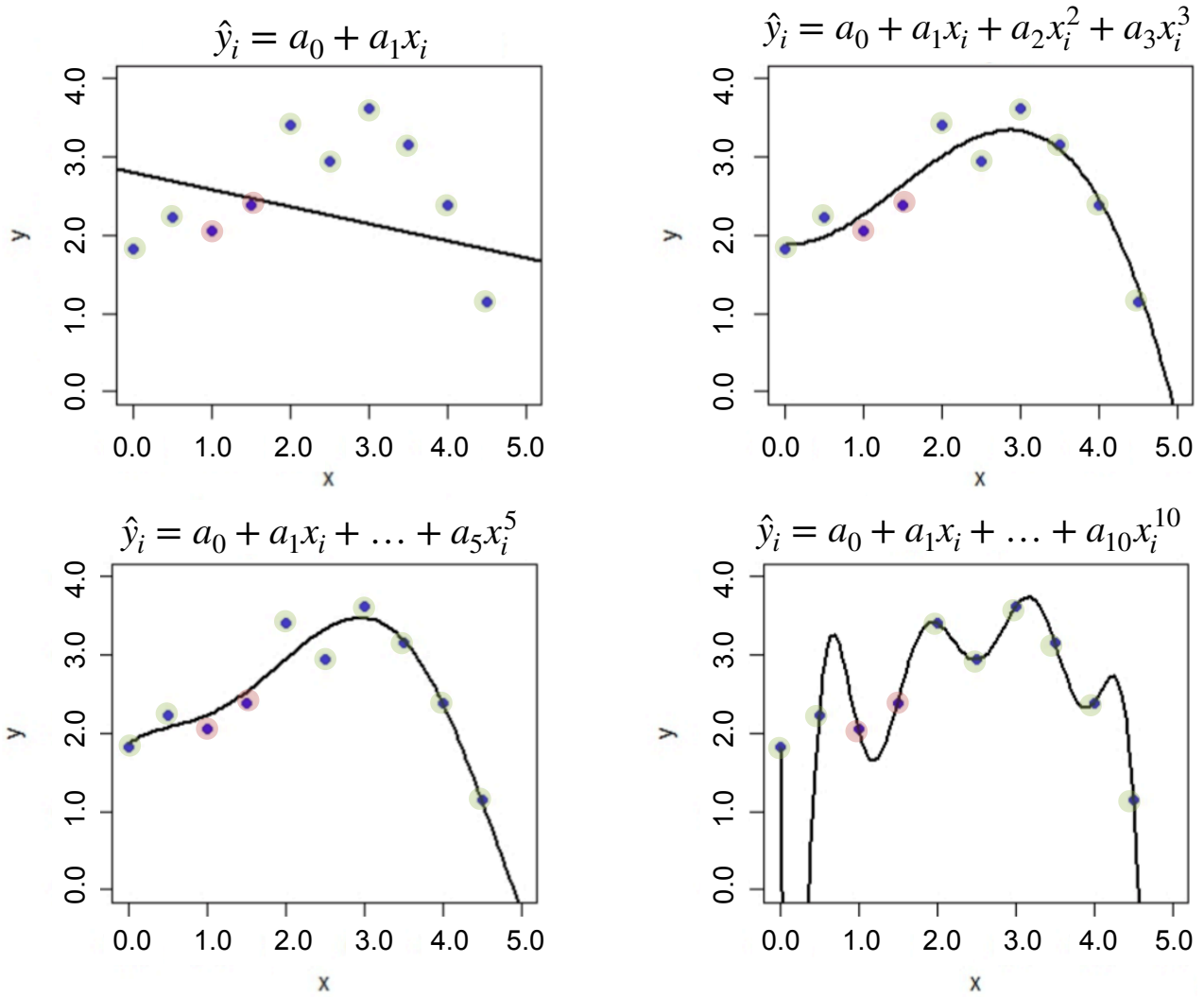
Compromis biais-variance



Causes de la nécessité de trouver un compromis entre biais et variance :

- La relation précise entre x et y est inconnue
- Les observations d'apprentissages $\{x_i, y_i\}_{i=1,\ldots,n}$ sont quasi systématiquement bruitées (ou la relation à capturer entre x et y est partielle).

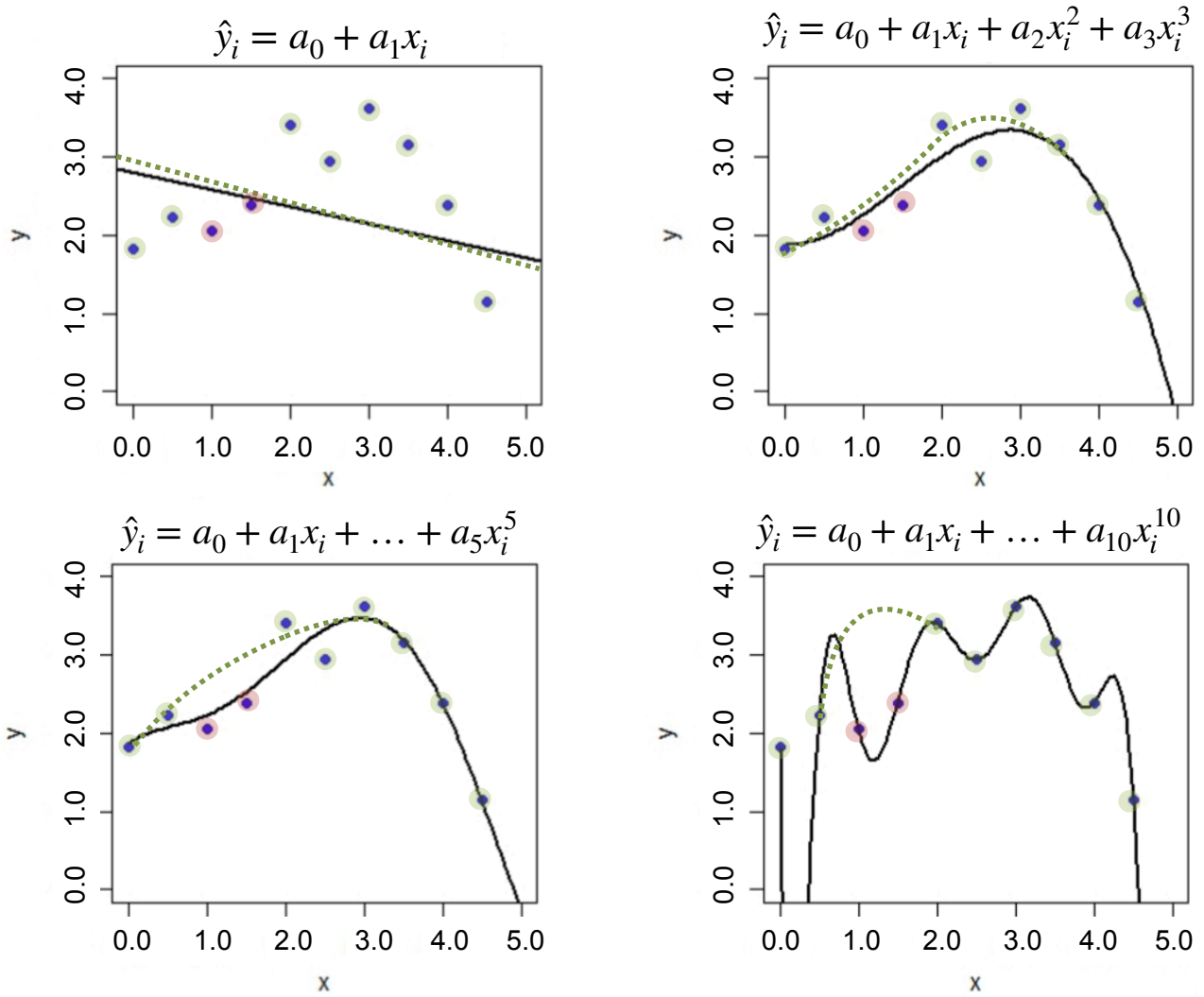
Validation croisée - Principe



Principe de la validation croisée :

- On distingue données d'apprentissage ● et de test ●
- On apprend sur les données d'apprentissage
- On mesure ensuite l'erreur sur les données test

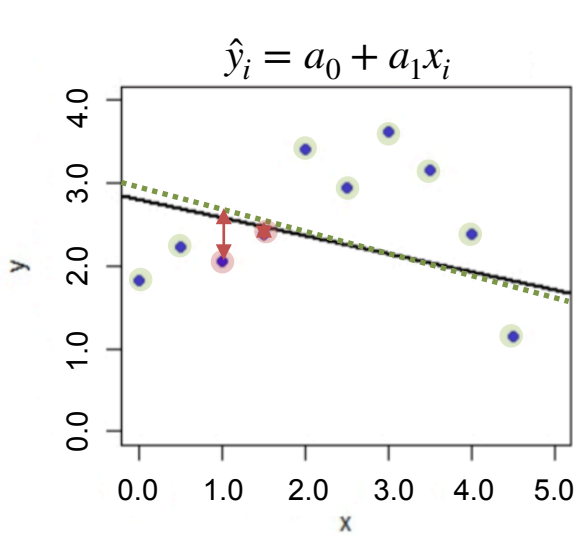
Validation croisée - Principe

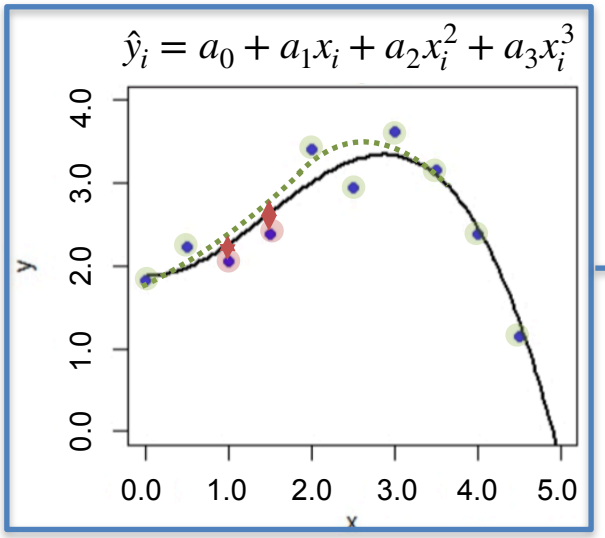


Principe de la validation croisée :

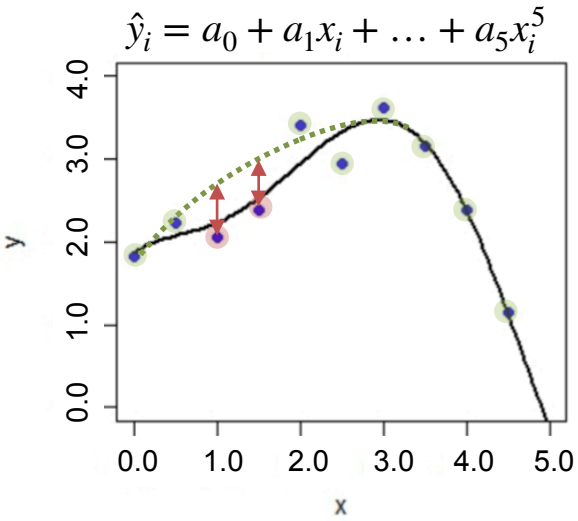
- On distingue données d'apprentissage ● et de test ●
- On apprend sur les données d'apprentissage
- On mesure ensuite l'erreur sur les données test

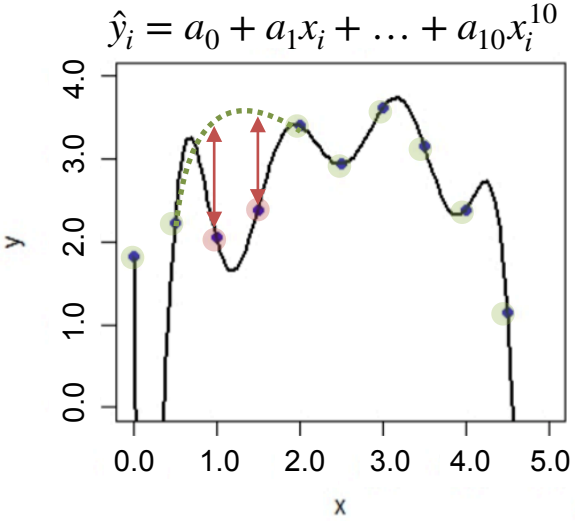
Validation croisée - Principe





Meilleur modèle pour les points considérés





<u>Principe de la validation croisée</u> :

- On distingue données d'apprentissage ● et de test ●
- On apprend sur les données d'apprentissage
- On mesure ensuite l'erreur sur les données test

Merci ☕