

# **Fondements statistiques de l'apprentissage automatique**

## **Chapitre 1 : Introduction**

Afin d'étudier quantitativement un phénomène à  $p \geq 1$  variables d'entrée  $X = (X^{(1)}, X^{(2)}, \dots, X^{(p)})$  et une variable de sortie  $Y$ , il est bien pratique de construire un modèle  $g$  qui explique par une relation mathématique les valeurs observées de  $Y$  en fonction des variables d'entrée :

$$Y = g(X^{(1)}, X^{(2)}, \dots, X^{(p)}) .$$

On distingue deux types de modèles :

1. *Modèles déterministes* : C'est une équation ou un ensemble d'équations qui émanent souvent de lois physiques, chimiques, économiques, ..., et représentent le comportement attendu du phénomène.
2. *Modèles statistiques* : Souvent, il est difficile de développer un modèle théorique car le phénomène étudié est trop complexe. On a alors recours à un modèle statistique basé non pas sur une théorie, mais sur des données observées.

**On connaît** les notes de  $n$  élèves au cours de l'année scolaire 2020-2021 ainsi que leur notes à un concours de fin d'année.

**On aimeraient prédire** les notes au concours des élèves de la promotion 2021-2022 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

Nouvel élève	Maths	Info	Français	Concours
	13	14	11	?

**On connaît** les notes de  $n$  élèves au cours de l'année scolaire 2020-2021 ainsi que leur notes à un concours de fin d'année.

**On aimeraient prédire** les notes au concours des élèves de la promotion 2021-2022 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  $i \in \{1, \dots, n\}$  durant l'année 2020-2021 :  $x_i = (x_i^1, x_i^2, \dots, x_i^p) \in \mathbb{R}^p$
- 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  $\mathbb{R}^p \rightarrow \mathbb{R}$   
Paramètres  $\Theta$  appris avec les  $(x_i, y_i)_{i=1, \dots, n}$

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

$$\widehat{y}_i = h_{\Theta}(x_i) = w_0 + \sum_{j=1}^p w_j x_i^j \quad \text{dont les paramètres sont } \Theta = \{w_0, w_1, \dots, w_p\}$$

- 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	?

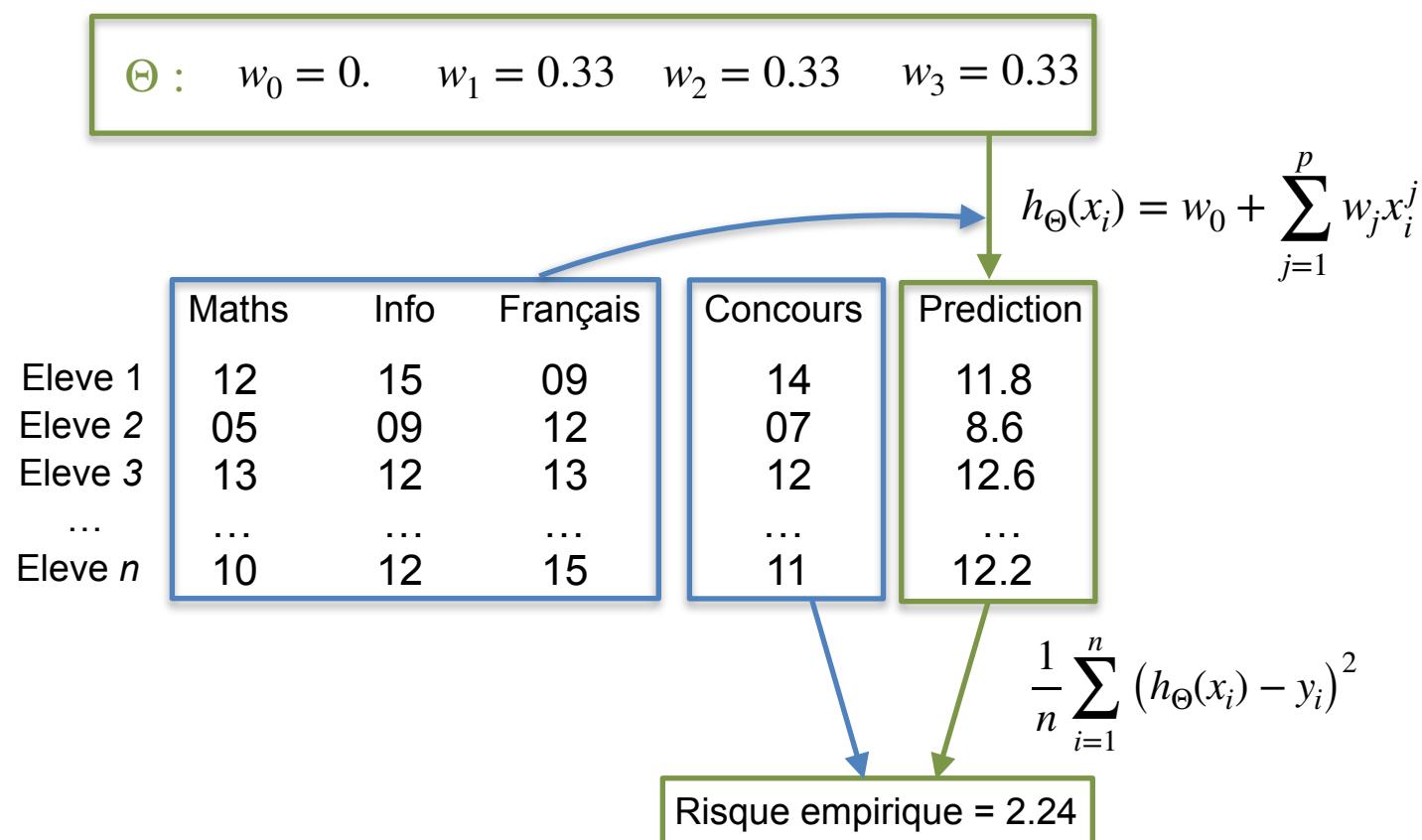
Prenons  $w_0 = 0.$ ,  $w_1 = 0.33$ ,  $w_2 = 0.33$  et  $w_3 = 0.33$  → Alors  $\widehat{y}_{new} = 12.54$

Les meilleurs paramètres  $\hat{\Theta}$  minimisent un **risque empirique** sur les  $(x_i, y_i)_{i=1, \dots, n}$ , par exemple :

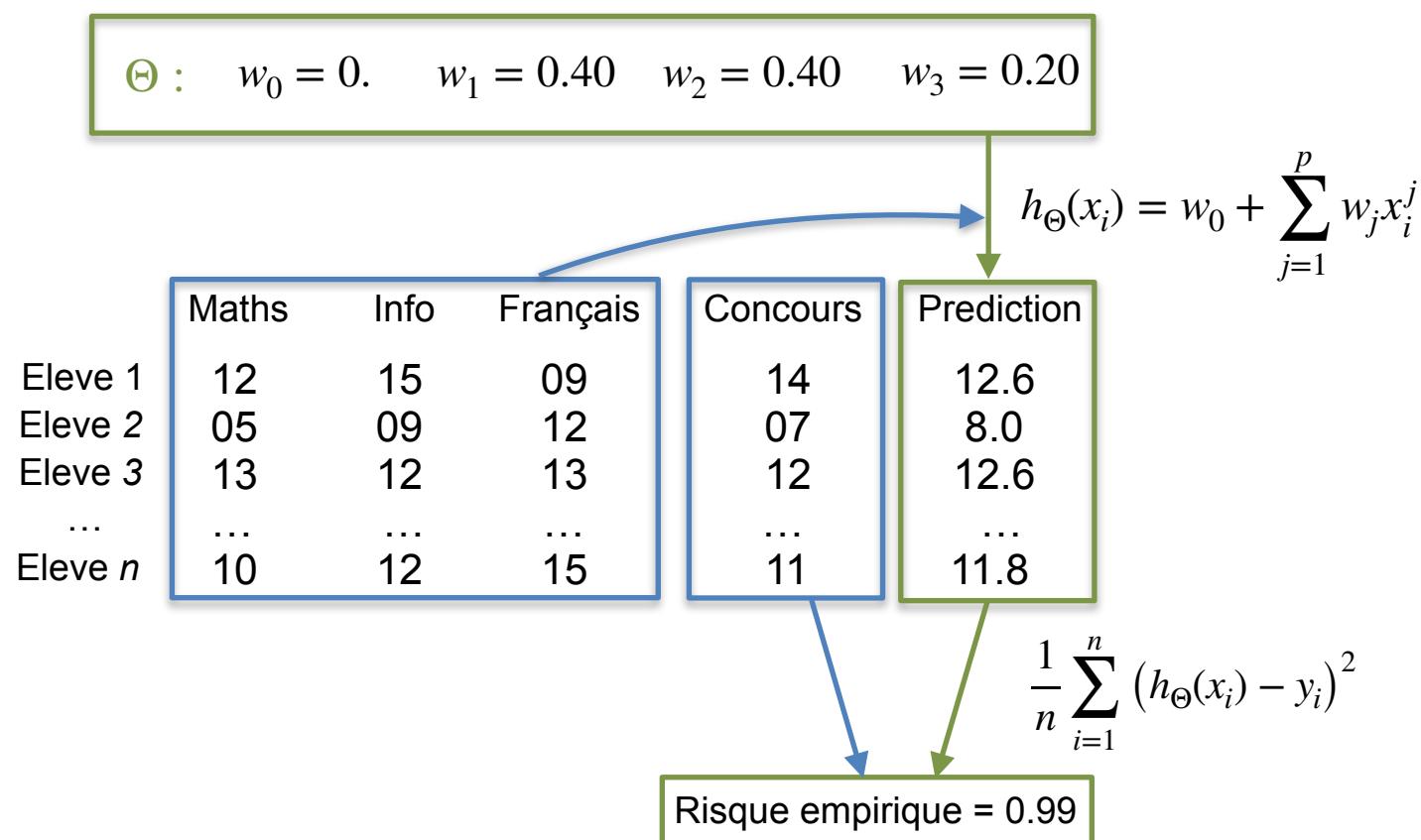
$$\begin{aligned}
 \hat{\Theta} &= \arg \min_{\Theta=\{w_0, \dots, w_p\}} \frac{1}{n} \sum_{i=1}^n \text{loss} (h_{\Theta}(x_i), y_i) \\
 &= \arg \min_{\Theta=\{w_0, \dots, w_p\}} \frac{1}{n} \sum_{i=1}^n (h_{\Theta}(x_i) - y_i)^2 \\
 &= \arg \min_{\Theta=\{w_0, \dots, w_p\}} \frac{1}{n} \sum_{i=1}^n \left( w_0 + \sum_{j=1}^p w_j x_i^j - y_i \right)^2
 \end{aligned}$$

Risque empirique  $R_{\Theta}((x_i, y_i)_{i=1, \dots, n})$

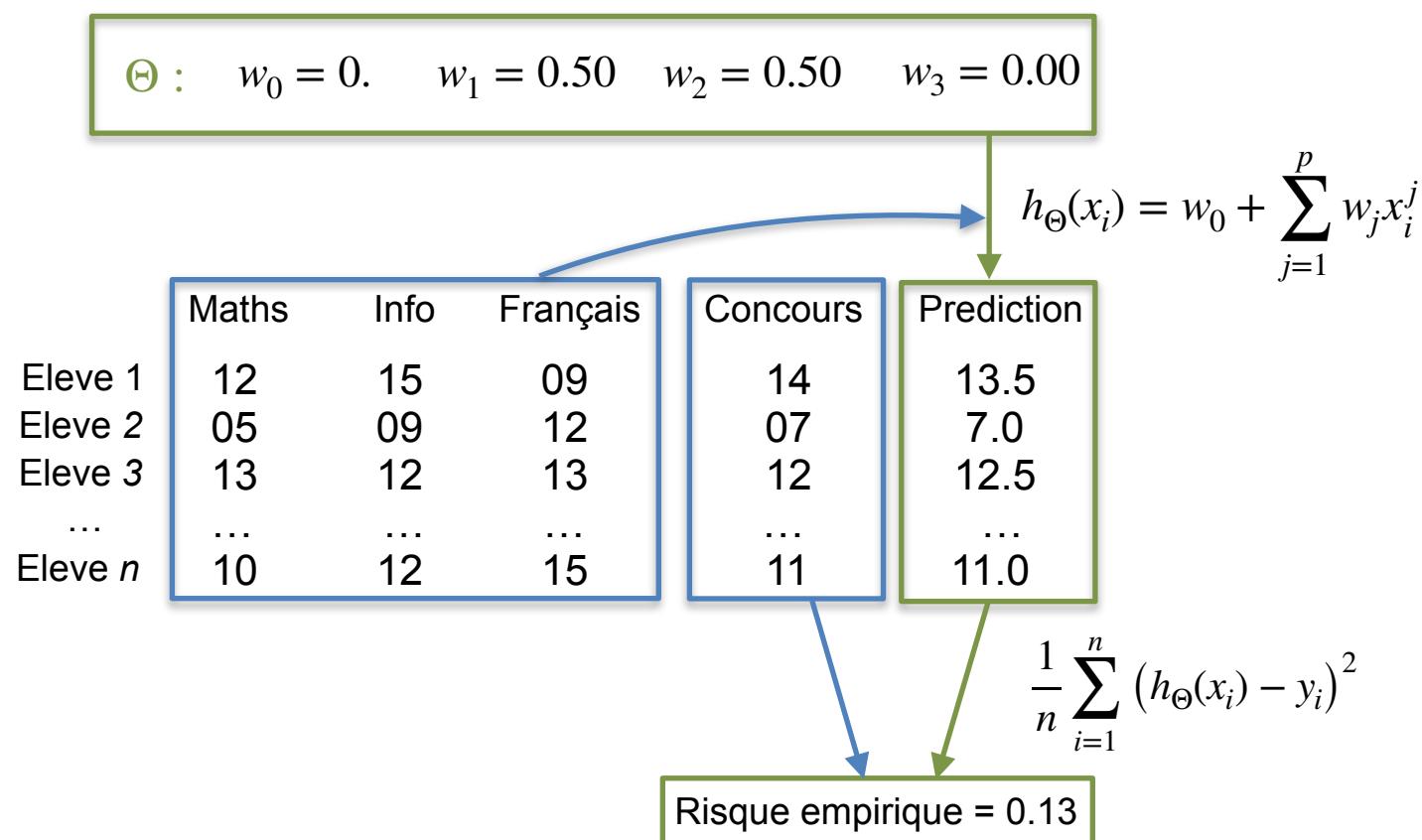
Les meilleurs paramètres  $\hat{\Theta}$  minimisent un **risque empirique** sur les  $(x_i, y_i)_{i=1,\dots,n}$ , par exemple :



Les meilleurs paramètres  $\hat{\Theta}$  minimisent un **risque empirique** sur les  $(x_i, y_i)_{i=1,\dots,n}$ , par exemple :



Les meilleurs paramètres  $\hat{\Theta}$  minimisent un **risque empirique** sur les  $(x_i, y_i)_{i=1,\dots,n}$ , par exemple :



**Coeur de ce cours :**

- Comprendre la modélisation *Statistique* de tels problèmes.
- Résoudre les problèmes en pratique (avec le modèle linéaire).
- Poser les bases de l'apprentissage automatique avec un oeil critique sur ce que représentent les données.
- Etendre et assimiler les modèles élémentaires.

**Objectif :**

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

## 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  $X_i = 0$ .
- 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  $s\sqrt{10}$  (voir appendice A).

**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 en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.

Exemples :

- Cas discret : Pile ou face dans un jeu, client Homme/Femme, patient jeune/adulte/âgé, ...
- Cas continu : Age d'un client, taille d'un patient, ...

**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 en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.

**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.

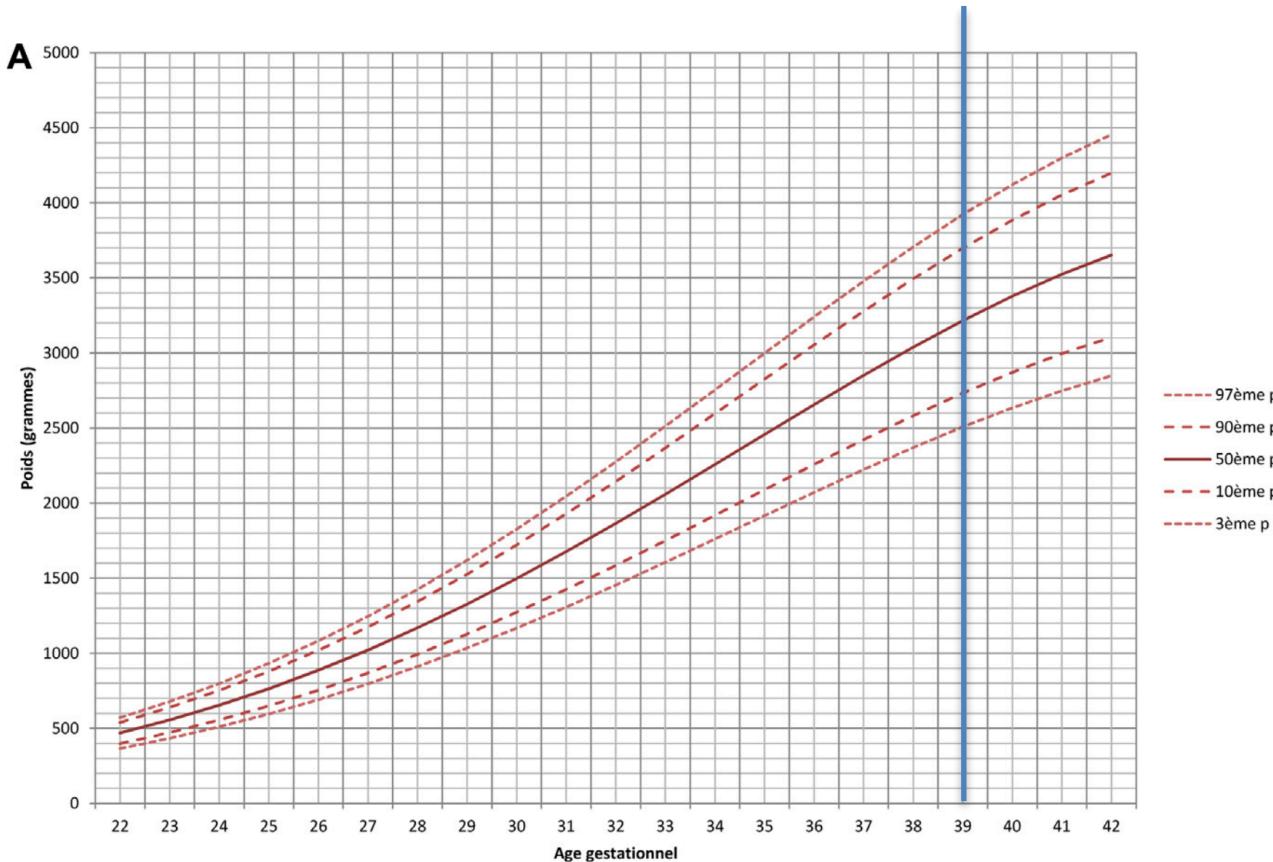
**Variable aléatoire** Une *variable aléatoire* (v.a.)  $X$  est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans  $\mathbb{R}$  ou un sous-ensemble de  $\mathbb{R}$ . On distinguera en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.

**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 en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.

Exemple :

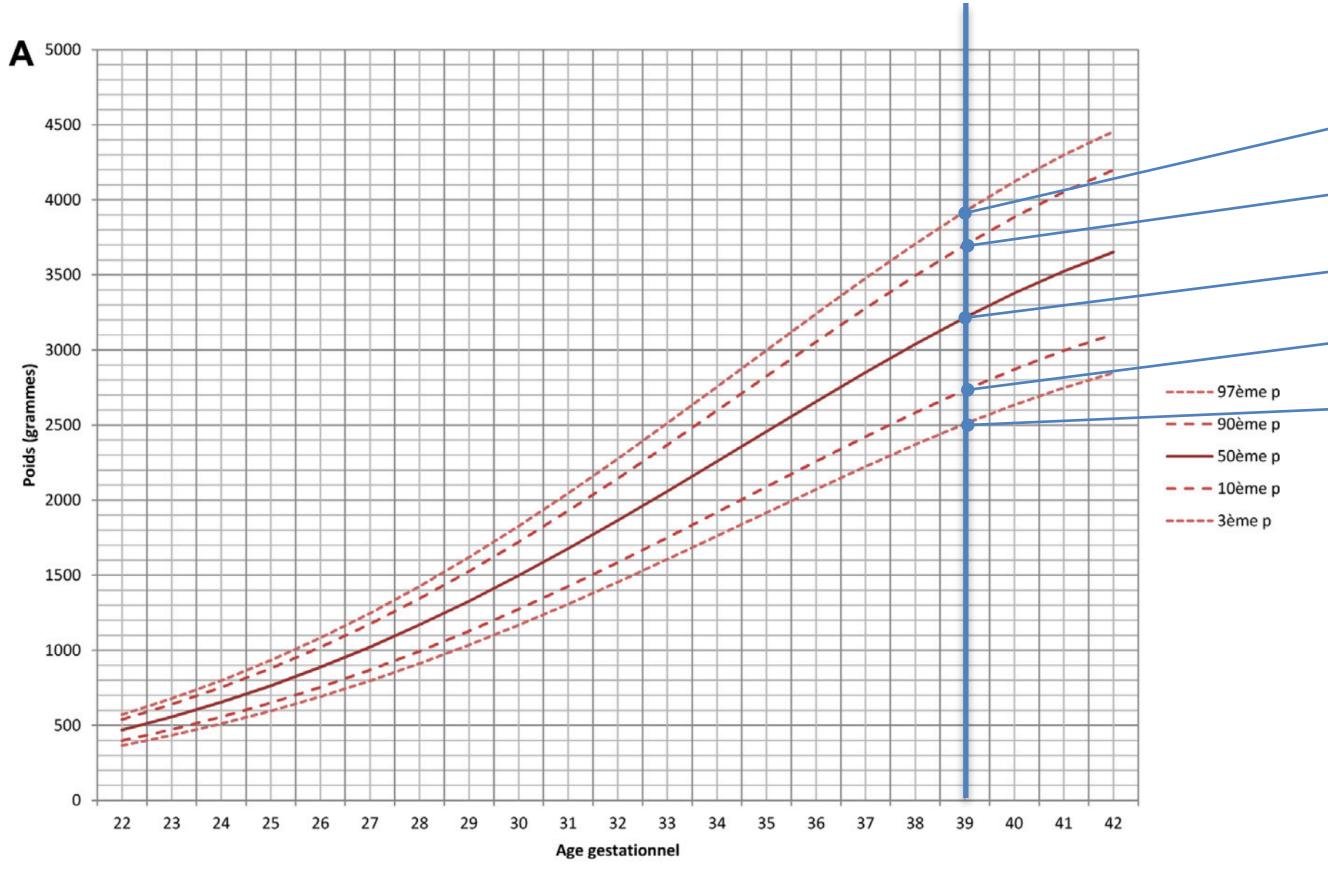
→ Fonction de répartition  $F_X(x) = \mathbb{P}(X < x)$  où  $X$  est le poids d'un enfant à 39 semaines



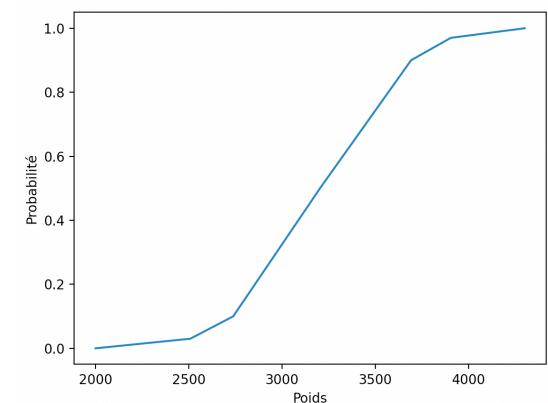
**Variable aléatoire** Une *variable aléatoire* (v.a.)  $X$  est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans  $\mathbb{R}$  ou un sous-ensemble de  $\mathbb{R}$ . On distinguera en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.

Exemple :

→ Fonction de répartition  $F_X(x) = \mathbb{P}(X < x)$  où  $X$  est le poids d'un enfant à 39 semaines



$$\begin{aligned}F_X(3904) &= 0.97 \\F_X(3691) &= 0.90 \\F_X(3203) &= 0.50 \\F_X(2738) &= 0.10 \\F_X(2508) &= 0.03\end{aligned}$$



**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 en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.

**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  $F_X(+\infty) = 1$ .

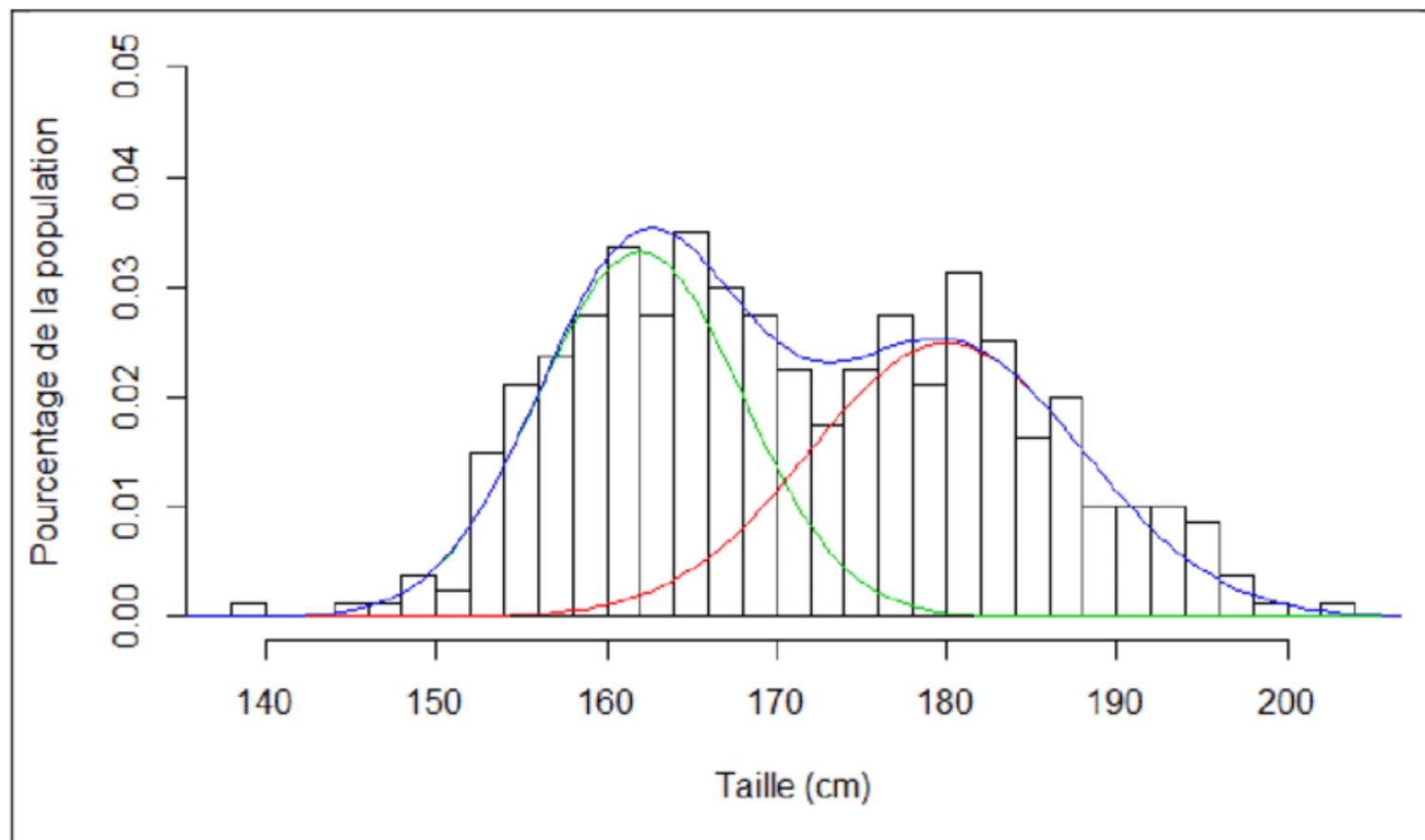
de répartition  $p_X(x)$ , la *densité de probabilité* pourra de même représenter la loi de probabilité d'une v.a.  $X$  suivant :

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

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 en particulier le *cas continu*, par exemple si  $X$  représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple  $X \in \{0, 1\}$  pour modéliser le résultat lorsque l'on joue à pile ou face.



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 probabilités/statistique, ainsi que de manipuler les concepts énoncés ci-dessus, il est intéressant de présenter maintenant le Théorème Central Limite (TCL).

Supposons que  $n$  variables aléatoires  $X_1, X_2, \dots, X_n$  indépendantes mais suivant une même loi de probabilité soient tirés. L'espérance (ou moyenne)  $m$  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  $s\sqrt{n}$ , *i.e.* :

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

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

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

Au delà de la connaissance du TCL lui même et de l'illustration des notions de la section 1.2.1, cet exemple nous amène un enseignement qui est (à mes yeux) l'essence de la modélisation statistique. En assemblant plusieurs variables aléatoires, nous avons créé un modèle aléatoire dont on peut étudier les propriétés statistiques telles que la moyenne mais aussi d'une certaine manière la précision/étendue/sensibilité. Ce type de modélisation se distingue alors de la modélisation déterministe qui ne s'intéresse qu'à l'équivalent de la moyenne ici.

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).

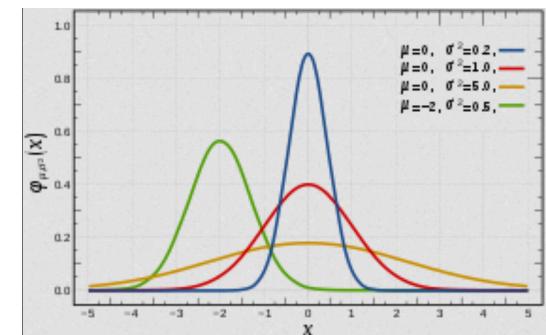
Voyons maintenant comment *estimer les paramètres d'un modèle* supposé représenter un phénomène observé à partir d'*observations* de ce phénomène.

Comme nous le verrons, cette technique est une généralisation de ce qui permet d'apprendre les paramètres d'un modèle d'apprentissage automatique (cf l'exemple sur la prédiction des notes au concours).

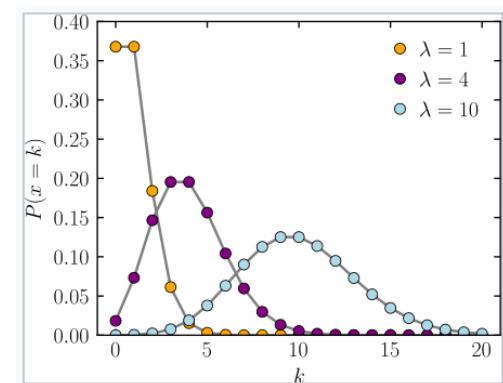
Exemple 1 : On suppose que la pièce lancée suit une loi de Bernoulli de paramètre  $p$ . Estimons  $p$  à partir de plusieurs lancés de pièces pour savoir si elle est bien équilibrée

$$\begin{aligned}\mathbb{P}(Pile) &= p \\ \mathbb{P}(Face) &= 1 - p\end{aligned}$$

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



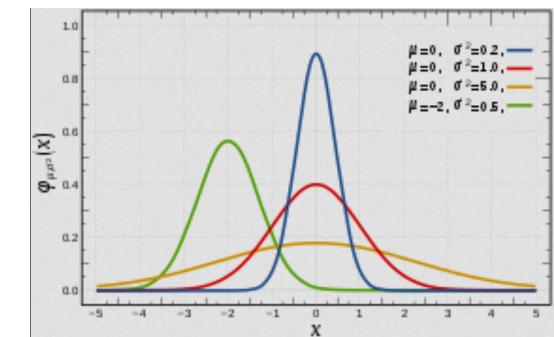
Exemple 3 : On suppose que le nombre de personnes devant nous à la caisse au magasin de journaux du quartier suit une loi de Poisson de paramètre  $\lambda$ . Estimons ce paramètre en observant sur quelques mois combien de personnes on avait devant nous.



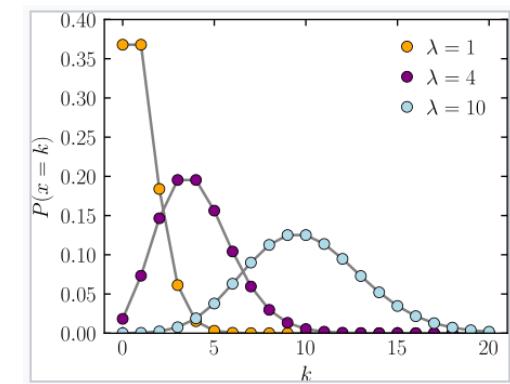
Exemple 1 : On suppose que la pièce lancée suit une loi de Bernoulli de paramètre  $p$ . Estimons  $p$  à partir de plusieurs lancés de pièces pour savoir si elle est bien équilibrée

$$\begin{aligned}\mathbb{P}(Pile) &= p \\ \mathbb{P}(Face) &= 1 - p\end{aligned}$$

Exemple 2 : On suppose que la taille des élèves de CM2 à Toulouse suit une loi normale de moyenne  $m$  et d'écart type  $\sigma$ . 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  $\lambda$ . Estimons ce paramètre en observant sur quelques mois combien de personnes on avait devant nous.



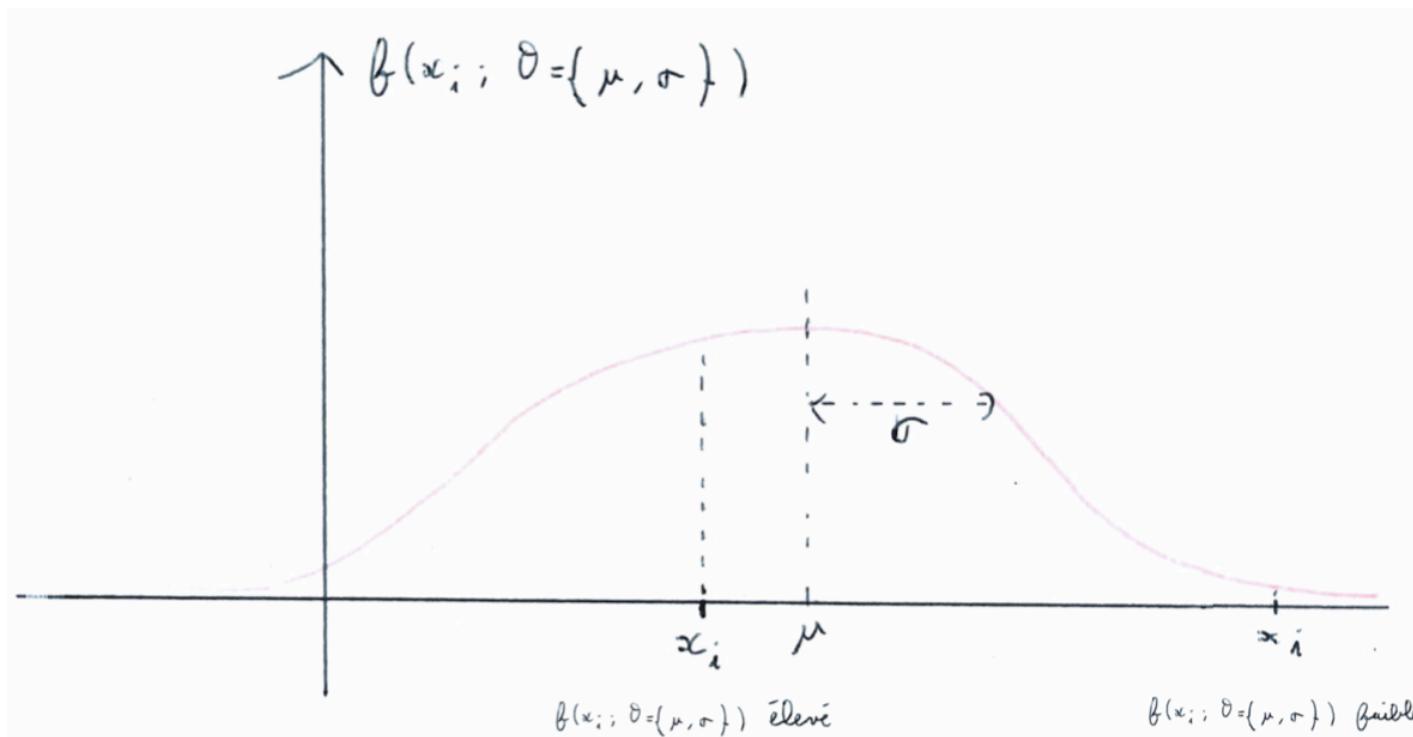
Pourquoi ces estimations ?

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

On dénote  $X$  une variable aléatoire (v.a.) supposée suivre une loi discrète (e.g. Bernoulli) ou continue (e.g. Normale) de paramètres  $\theta$ . On note aussi  $x_1, \dots, x_i, \dots, x_n$  les observations de  $X$ .

Pour une observation  $x_i$  donnée, on modélise alors la loi de  $X$  avec la fonction  $f(x_i; \theta)$ . Cette fonction vaut  $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$  si  $X$  est une v.a. discrète et  $f(x_i; \theta) = f_\theta(x_i)$  si  $X$  est continue, où  $f_\theta(x_i)$  est la densité de la loi en fonction de ses paramètres  $\theta$ .

Pour des paramètres  $\theta$  donnés (ex : moyenne et écart type d'une loi normale),  $f(x_i; \theta)$  sera alors d'autant plus élevée que  $x_i$  a des chances d'être tirée en fonction des  $\theta$ .



La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

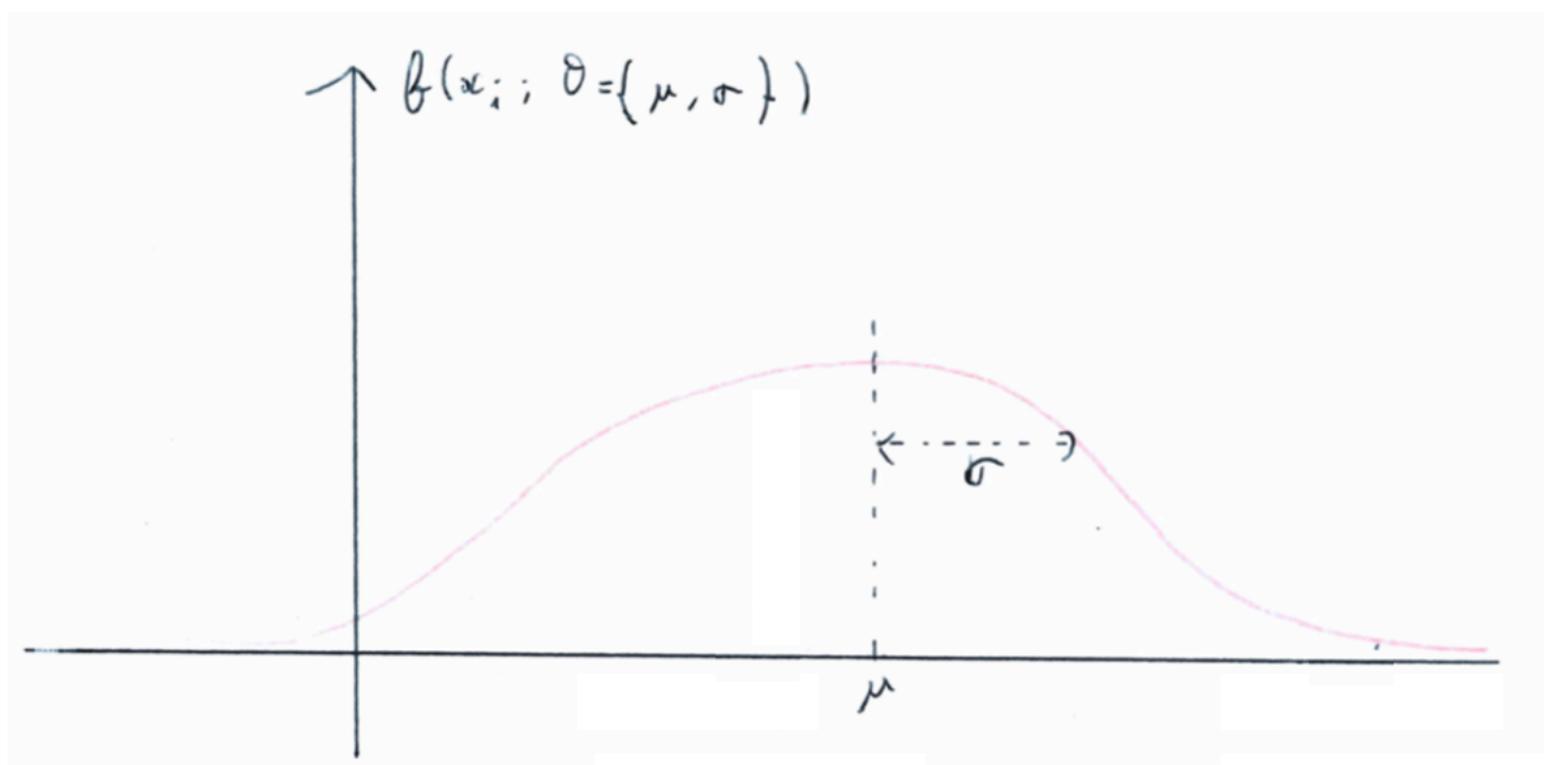
$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

Dans l'exemple de pile ou face, supposons que l'on souhaite vérifier empiriquement si une pièce est équilibrée ou non. On modélisera  $\mathbb{P}(X = 1) = f(X_i = 1; \theta = \{p\}) = p$  et  $\mathbb{P}(X = 0) = f(X_i = 0; \theta = \{p\}) = 1 - p$ , puis on réalisera  $n$  observations de  $X$  en tirant à pile ou face. La vraisemblance sera alors  $L(\theta = \{p\}) = \prod_{i=1}^n (1_{X_i=1}p + 1_{X_i=0}(1 - p))$ .

Supposons que sur  $n = 10$  tirages, on observe 4 'piles' et 6 'faces'. En simplifiant légèrement les notations, la vraisemblance du paramètre  $p$  par rapport à notre modèle et nos observations empiriques sera alors  $L(p) = p^4(1 - p)^6$ . 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,  $p = 0.5$  semble le plus vraisemblable.

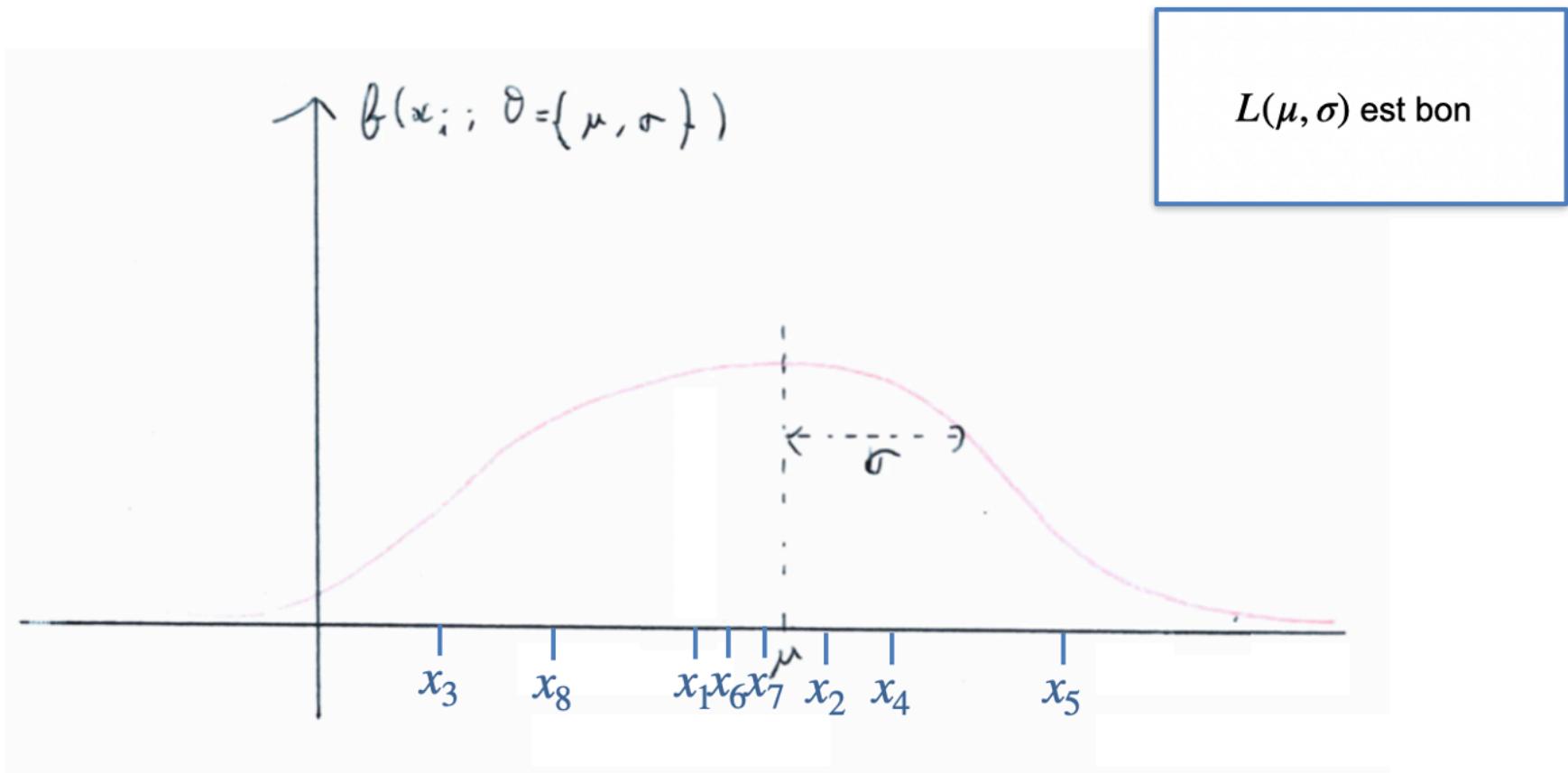
La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



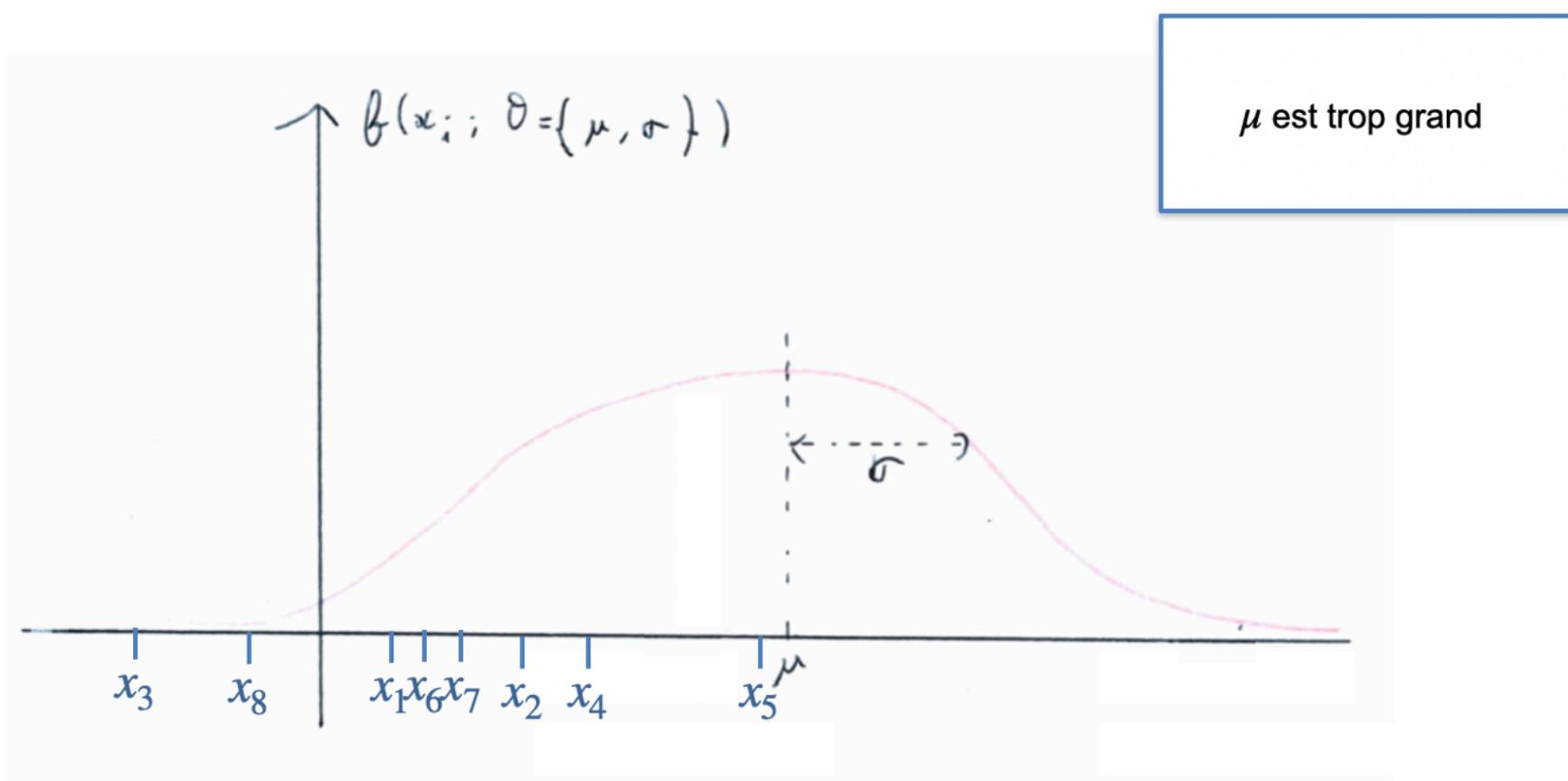
La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



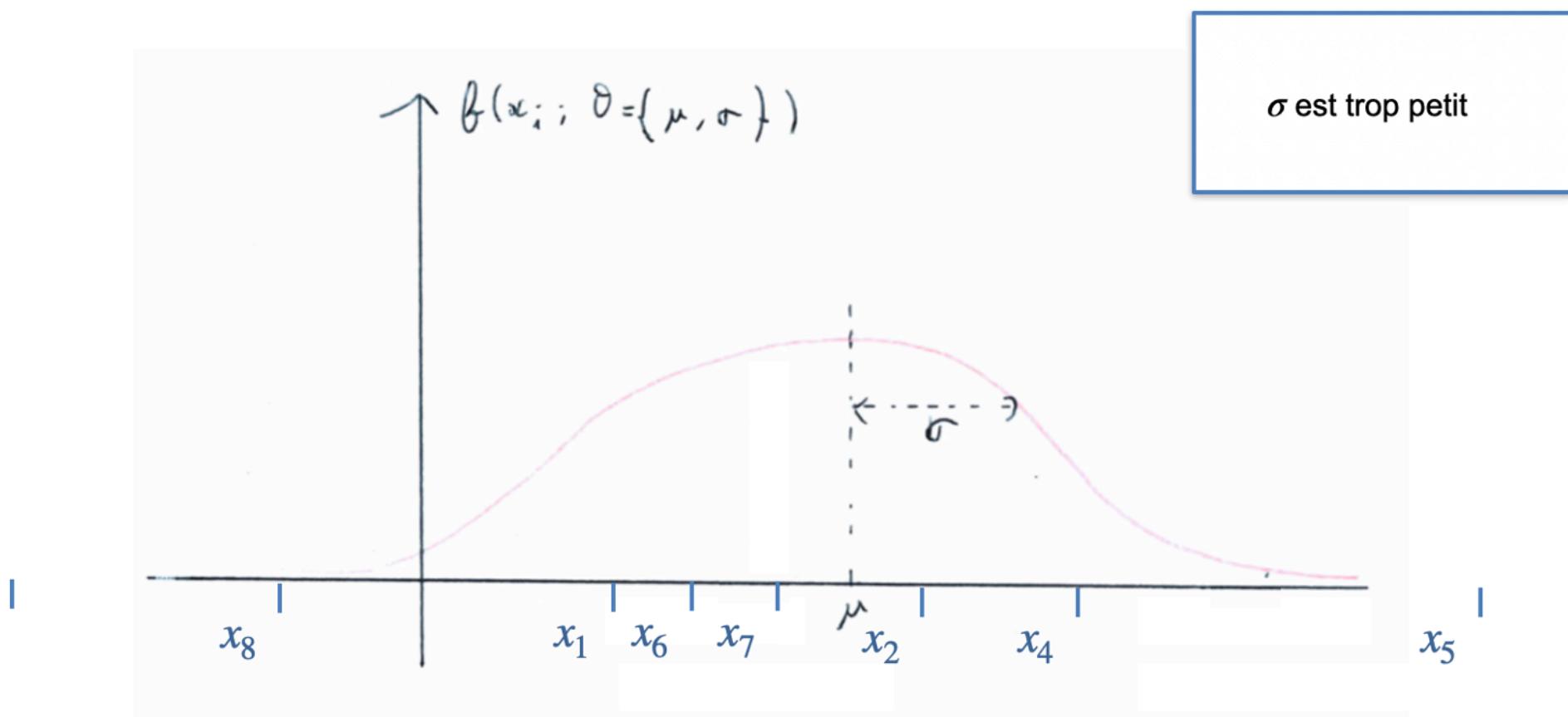
La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



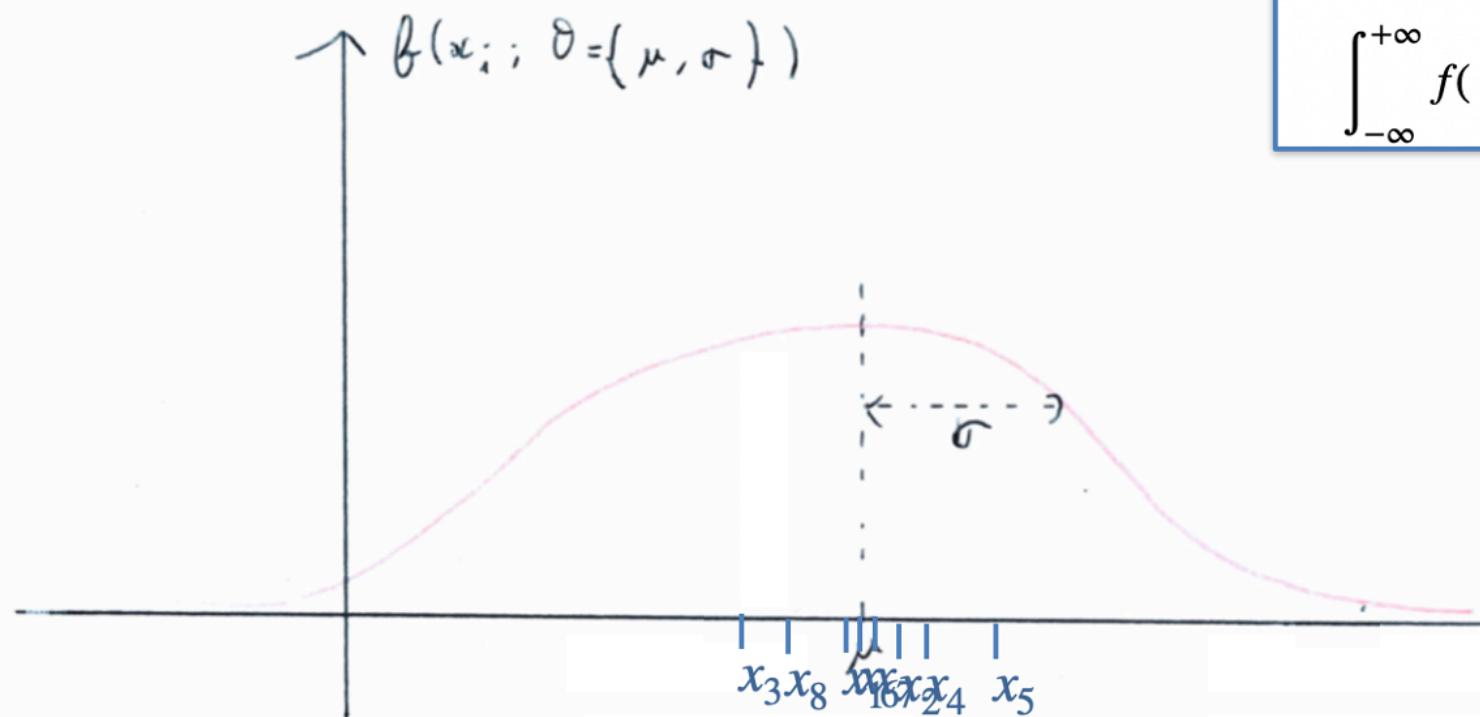
La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



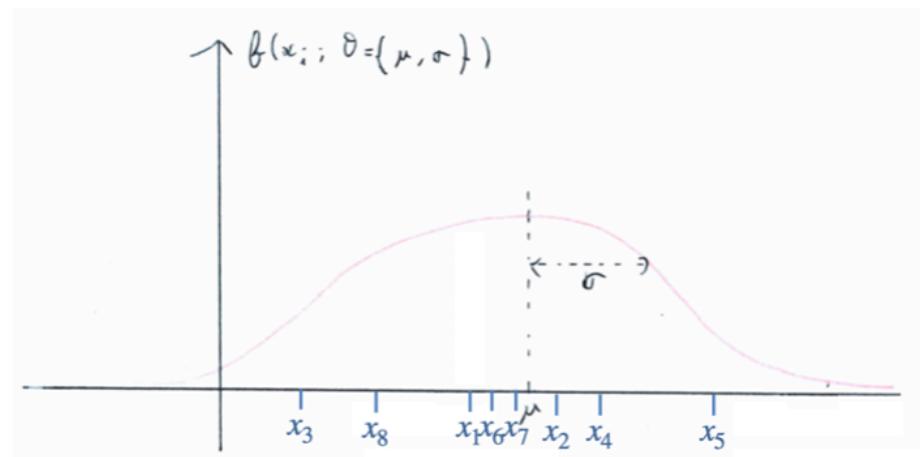
La vraisemblance des paramètres  $\theta$  en fonction des observations  $x_1, \dots, x_n$  est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

Afin de trouver les paramètres d'une loi les plus vraisemblables, une fois les  $\{x_i\}_{i=1,\dots,n}$  connus, on calculera le *maximum de vraisemblance* :

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

qui renverra les paramètres les plus vraisemblables en fonction des observations et de la loi choisie.



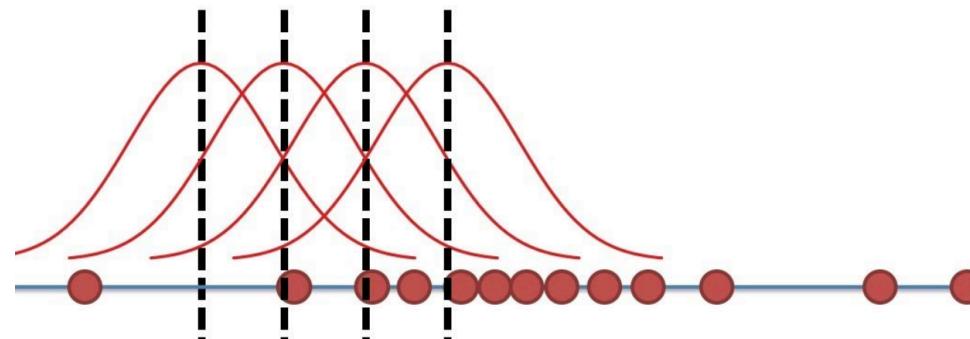
De manière générale, on calculera le maximum de vraisemblance :

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

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

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

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



Retournons maintenant à notre exemple ‘notes au concours’ avec un regard de ‘statisticien’ :

Les meilleurs paramètres  $\hat{\Theta}$  minimisent un **risque empirique** sur les  $(x_i, y_i)_{i=1,\dots,n}$ , par exemple :

$$\hat{\Theta} : \boxed{w_0 = 0.0} \quad w_1 = 0.40 \quad w_2 = 0.40 \quad w_3 = 0.20$$

$$h_{\hat{\Theta}}(x_i) = w_0 + \sum_{j=1}^p w_j x_i^j$$

	Maths	Info	Français	Concours	Prediction
Eleve 1	12	15	09	14	12.6
Eleve 2	05	09	12	07	8.0
Eleve 3	13	12	13	12	12.6
...	...	...	...	...	...
Eleve $n$	10	12	15	11	11.8

$$\frac{1}{n} \sum_{i=1}^n (h_{\hat{\Theta}}(x_i) - y_i)^2$$

Risque empirique = 0.99

Utilisons une modélisation qui plus flexible sur l'erreur

Faisons un point maintenant... 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  $\Theta$  sont notre inconnue
- On considère l'erreur de prédiction :  $e_{i,\Theta} = h_{\Theta}(x_i) - y_i$

Faisons un point maintenant... 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  $\Theta$  sont notre inconnue
- On considère l'erreur de prédiction :  $e_{i,\Theta} = h_{\Theta}(x_i) - y_i$

On fait l'hypothèse :  $e_i \sim \mathcal{N}(0, \sigma)$

et on maximise la vraisemblance :  $L(\Theta) = \prod_{i=1}^n f(x_i, y_i; \Theta, \sigma)$

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

On suppose disposer d'observations  $\{y_i\}_{i=\{1,\dots,n\}}$  que l'on souhaite prédire/deviner à partir de observations correspondantes  $\{x_i\}_{i=\{1,\dots,n\}}$ , où chaque  $y_i$  correspond à  $x_i$  (voir l'exemple introductif par exemple). Dans ce cours, et très souvent en apprentissage automatique, on va alors optimiser les paramètres  $\theta$  d'un modèle  $f_\theta$  pour prédire au mieux les  $y_i$  avec  $\hat{y}_i = f_\theta(x_i)$ .

Faisons l'hypothèse que les erreurs d'approximation du modèle  $e_i = y_i - f_\theta(x_i)$  suivent une loi normale centrée, *i.e.*  $e_i \sim \mathcal{N}(0, \sigma)$ . Ce choix par défaut est commun et semble raisonnable quand  $f_\theta$  est bien calibré. Nous pouvons alors utiliser le principe de maximum de vraisemblance pour estimer les paramètres  $\theta$  du modèle  $f_\theta$ .

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

On suppose disposer d'observations  $\{y_i\}_{i=\{1,\dots,n\}}$  que l'on souhaite prédire/deviner à partir de observations correspondantes  $\{x_i\}_{i=\{1,\dots,n\}}$ , où chaque  $y_i$  correspond à  $x_i$  (voir l'exemple introductif par exemple). Dans ce cours, et très souvent en apprentissage automatique, on va alors optimiser les paramètres  $\theta$  d'un modèle  $f_\theta$  pour prédire au mieux les  $y_i$  avec  $\hat{y}_i = f_\theta(x_i)$ .

Faisons l'hypothèse que les erreurs d'approximation du modèle  $e_i = y_i - f_\theta(x_i)$  suivent une loi normale centrée, *i.e.*  $e_i \sim \mathcal{N}(0, \sigma)$ . Ce choix par défaut est commun et semble raisonnable quand  $f_\theta$  est bien calibré. Nous pouvons alors utiliser le principe de maximum de vraisemblance pour estimer les paramètres  $\theta$  du modèle  $f_\theta$ .

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

Cette technique d'estimation est celle dite au sens des moindres carrés. Nous la retrouvons très couramment en apprentissage automatique et son interprétation est particulièrement intuitive. Elle doit notamment sa popularité au fait qu'il est aisément de calculer son gradient par rapport aux paramètres  $\theta$  si on sait calculer le gradient de  $f_\theta$  par rapport à  $\theta$  :

$$\nabla_{\theta} e_i^2 = 2(y_i - f_{\theta}(x_i)) \nabla_{\theta} f_{\theta}(x_i)$$

Cela ouvre la porte aux techniques d'optimisation par descente de gradient qui sont quasi systématiques en apprentissage automatique.

Pour un public avisé, il faudra se souvenir du fait que la pertinence de l'estimation de paramètres d'un modèle au sens des moindres carrés repose sur une hypothèse de normalité de l'erreur.