APM-41033-EP : Introduction aux méthodes statistiques

Cours 1 — Modèles statistiques

Aymeric Dieuleveut

École polytechnique

Septembre 2026

Qu'est-ce que la statistique ?

La statistique est l'étude de la collecte, de la modélisation et du traitement des données.

Médecine, Santé

- épidémiologie
- génomique, protéomique
- médecine de précision

Sciences de l'environnement

- pollution
- météorologie
- climat

Assurance, finance

- Finance (Ruppert & Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer, 2015) : modélisation des actifs (actions, options, indices, futures).
- Assurance et actuariat : calcul des risques, primes, défaut.

Apprentissage statistique & traitement du signal

- Reconnaissance de parole
- Vision par ordinateur
- Traduction automatique

Marketing

- Système de recommandation
- Optimisation séquentielle (sites web, etc.)
- Scoring clients

⚠ Tous ces exemples sont des exemples avancés, dans lesquels les données observables sont complexes

Des questions plus simples pour commencer

1. Quelle est la probabilité qu'une pièce retombe sur "pile" ?
   Les pièces de "1ct" sont-elles biaisées ?

2. Quelle est la taille moyenne de quelqu'un dans cette salle ?
   Les élèves de Polytechnique sont-ils plus grands que la moyenne ?

3. Quelle proportion des 2A de Polytechnique a déjà utilisé un modèle d'IA pour faire une partie d'un DM alors que c'était interdit ?

Apporter une réponse éventuelle à ces questions demande :

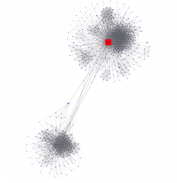
- De les reformuler dans un langage mathématique
- De collecter, modéliser et analyser des données.

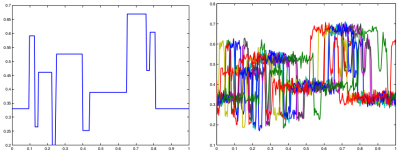
Formalisme statistique - données

Point de départ : des données

$$(x_1, \ldots, x_n).$$

- numériques (scalaires, vectorielles, matricielles) : $x_i \in \mathbb{R}^d$ ou $x_i \in \mathbb{R}^{p \times q}$
- symboliques (labels, catégorielles) : $x_i \in \{0, 1\}$ (par exemple)
- mais aussi des données plus structurées : graphes, des séries chronologiques, des fonctions (courbes), des textes (LLM), des images, etc.





(a) (b)

Exemples :

- Résultats de multiples lancers de pile ou face
- Taille des personnes présentes dans l'amphi

Comment formaliser les questions ?

Modélisation statistique

Nous allons *modéliser* les observations par des réalisations de variables aléatoires :

$\rightarrow$ Données = réalisation d'un élément aléatoire

$$Z = (X_1, \ldots, X_n) \ , \quad Z(\omega) = (X_1(\omega), \ldots, X_n(\omega)) = (x_1, \ldots, x_n) \ .$$

avec $(X_1, \ldots, X_n)$ des variables aléatoires sur un espace $(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow (\mathsf{Z}, \mathcal{Z})$, l'espace des observations, avec $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé (toujours omis)

Nous allons *modéliser* la loi de ces variables aléatoires :

$\rightarrow$ la loi du vecteur aléatoire

$$Z = (X_1, \ldots, X_n)$$

est partiellement connue : c'est un élément d'une famille $\mathcal{C}$ de lois.

Définition : Modèle statistique

Un modèle statistique est un triplet $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ avec : un espace mesurable $(\mathsf{Z}, \mathcal{Z})$, dit l'espace des observations, et une famille de probabilités $\mathcal{C}$ sur $(\mathsf{Z}, \mathcal{Z})$.

On pourra éventuellement appeler la famille de lois $\mathcal{C}$ le modèle statistique.

Exemples

J'observe le nombre de "piles" issu de n lancers d'une pièce.

1. J'observe une unique réalisation d'une variable aléatoire à valeurs dans $\mathbb{N}$
2. On choisit la tribu des parties $\mathcal{P}(\mathbb{N})$
3. On suppose que les lancers sont indépendants, de loi de Bernoulli $\mathcal{B}\mathrm{er}(p)$ pour un $p \in [0,1]$ (on note p le paramètre d'une loi de Bernoulli) : notre observation a donc pour loi $\mathrm{Bin}(n,p)$.

Mon modèle est

$$\Big(\mathbb{N}, \mathcal{P}(\mathbb{N}), \{\mathrm{Bin}(n,p), p \in [0,1]\}\Big)$$

J'observe la taille de n polytechnicien(ne)s.

1. J'observe n valeurs réelles, mon observation est dans $\mathbb{R}^n$
2. J'utilise la tribu borélienne $\mathcal{B}(\mathbb{R}^n)$
3. Je modélise la taille par une loi gaussienne de moyenne $\mu \geq 0$ et variance σ^2, et modélise les différentes observations par des v.a. indépendantes.

Mon modèle est

$$\Big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathrm{I}_n), \mu \geq 0, \sigma^2 \geq 0\}\Big)$$

Convention du cours : $\sigma^2 = 0$ est admis (masse de Dirac en μ) ; la densité gaussienne n'existe que pour $\sigma^2 > 0$.

Définition : Modèle paramétrique

Le modèle est dit paramétrique s'il s'écrit $\mathcal{C} = \{\mathbb{P}_\theta \, : \, \theta \in \Theta\}$ avec $\Theta \subseteq \mathbb{R}^d$, d fini, pour une paramétrisation naturelle. Il est dit non paramétrique lorsqu'il ne se décrit pas par un nombre fini de paramètres réels (par exemple : toutes les lois à densité continue sur $\mathbb{R}$).

Objectifs de la statistique

À partir des observations, on va chercher à affiner notre connaissance de la loi de l'observation.

Plus spécifiquement, on distingue plusieurs tâches :

- estimation ponctuelle : donner une valeur plausible du paramètre θ,
- estimation par région : déterminer un sous-ensemble $\Theta_0(Z) \subset \Theta$ auquel le paramètre θ appartient de façon plausible
- test : décider si le paramètre θ appartient à un sous-ensemble Θ_0 et évaluer la plausibilité de la décision.
- prédire : donner une valeur plausible pour une quantité non observée, par exemple prédire la valeur d'une sortie Y à partir d'une entrée X

Remarque : Statistique et Probabilité

- Probabilité : les lois sont supposées connues... Étant donnée une loi de probabilité $\mathbb{P}$ sur un espace mesurable $(\Omega, \mathcal{A})$ et un vecteur aléatoire Z, l'objet du calcul des probabilités est d'évaluer des quantités de la forme

$$\mathbb{P}(f(Z) \geq c), \quad \mathbb{E}[f(Z)], \text{etc.}$$

- Statistiques : on cherche à résoudre un problème inverse. Étant donnée une réalisation $z = (x_1, \ldots, x_n)$ d'un élément aléatoire $Z = (X_1, \ldots, X_n)$, on cherche à inférer certaines caractéristiques de la loi de ce vecteur aléatoire.

Discussion sur le rôle de la modélisation

George Box, **1919-2013**

💡 Un modèle est une hypothèse !
C'est même précisément notre hypothèse de départ pour toute la partie mathématique de la statistique.

1. *Tous les modèles sont faux, mais certains sont utiles !* George Box

En pratique la "pertinence" d'une modélisation statistique va dépendre du contexte.

- Modéliser le résultat d'un pile ou face par une $\mathcal{B}\mathrm{er}(p)$.
- Modéliser la taille des personnes par une $\mathcal{N}(\mu, \sigma^2)$
- Modéliser les réponses d'un sondage téléphonique par des Bernoulli i.i.d.
- Modéliser le temps d'attente entre deux clients dans une file par une loi exponentielle, etc.

→ ⚠ Un modèle statistique est une **description mathématique simplifiée** d'un phénomène (aléatoire) réel.
En choisissant un modèle, on fixe un cadre théorique dans lequel on pourra raisonner, estimer des grandeurs inconnues, prédire de nouvelles observations ou tester des hypothèses. Comme toute simplification, un modèle ne capture pas toute la réalité. Il faut donc prendre la "conclusion" statistique dans le monde réel avec précaution (tout comme la conclusion d'un modèle physique, également "inexact").

Remarque : la capacité prédictive du monde des mathématiques est une chose extraordinaire. [1]
C'est une des choses fascinantes des mathématiques appliquées. ♡

1. The Unreasonable Effectiveness of Mathematics in the Natural Sciences

En statistique mathématique, le modèle comme point de départ

Dans ce cours, on se focalisera sur la *statistique mathématique*. On prendra le modèle comme point de départ.

$\rightarrow$ Tout ce qu'on écrit à partir de là est **mathématiquement exact.**

On considère le modèle $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ veut dire "on suppose observer une réalisation d'une variable aléatoire Z à valeurs dans Z dont la loi est un élément de $\mathcal{C}$."

Par exemple *On considère le modèle* $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\lambda), \lambda > 0\})$.
$\Leftrightarrow$ On suppose observer une variable aléatoire $X \sim \mathcal{E}(\lambda)$ pour un $\lambda > 0$.
Dans le cours, on soulignera cette hypothèse et la dépendance en le paramètre via la notation $\mathbb{E}_\lambda[X], \mathbb{P}_\lambda(X \in A)$

$\mathbb{E}_\lambda[X] :=$ espérance de X sous l'hypothèse que $X \sim \mathcal{E}(\lambda)$.

Un premier résumé !

1. Le modèle statistique est le **point de départ fondamental de l'approche statistique**.
2. (c'est pour ça qu'on lui consacre un cours !)

3. Tout ce qui en est déduit est rigoureux
4. Comme toute modélisation, elle peut être inexacte

5. Le modèle s'écrit comme un triplet $(\mathsf{Z}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$.
6. Il permet de formaliser les questions statistiques que nous allons aborder : estimation / test / région de confiance / etc.

Next :

1. Propriétés des modèles, notion de statistique, de modèle induit, n-échantillon, etc.
2. Applications
3. Organisation générale du cours.

Définition : Identifiabilité

Soit un modèle $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$, avec $\mathcal{C} = \{\mathbb{P}_\theta, \theta \in \Theta\}$.
Le modèle est identifiable si la fonction $\theta \mapsto \mathbb{P}_\theta$ est injective, i.e. si $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ implique $\theta = \theta'$.

$\rightarrow$ Cela signifie que la connaissance complète de la loi $\mathbb{P}_\theta$ permet d'identifier de façon unique θ.
$\rightarrow$ C'est donc une propriété "minimale" pour espérer retrouver θ à partir des observations.

Exemples :

- Tous les modèles vus ci-dessus sont identifiables.

Identifiabilité : un modèle non identifiable

Exemple : Mélange gaussien

Considérons un modèle de mélange gaussien :

$$\mathcal{C} = \left\{ p_{\pi,\mu_1,\mu_2} \cdot \mathrm{Leb}, \quad \pi \in [0,1]\,,\ (\mu_1,\mu_2) \in \mathbb{R}^2 \right\}$$

où :

$$p_{\pi,\mu_1,\mu_2} = \pi p_{\mathcal{N}(\mu_1,\sigma^2)} + (1-\pi) p_{\mathcal{N}(\mu_2,\sigma^2)}$$

- $\pi \in [0,1]$ est la proportion de la première composante,
- μ_1, μ_2 sont les moyennes,
- σ^2 est la variance *connue*.

N'est pas identifiable : $p_{\pi,\mu_1,\mu_2} = p_{1-\pi,\mu_2,\mu_1}$

(d'ailleurs, à quel contexte serait-il approprié ?)  Comment rendre le modèle ci-dessus identifiable ?

Définition : Statistique

Soient $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ un modèle statistique et $(\mathsf{T}, \mathcal{T})$ un espace mesurable.
On appelle statistique sur $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ une application mesurable T de $(\mathsf{Z}, \mathcal{Z})$ à valeurs dans $(\mathsf{T}, \mathcal{T})$.

Exemple

Si $(\mathsf{Z}, \mathcal{Z}, \mathcal{C}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_\theta, \theta \in \Theta\})$,

1. La i-ème observation est la statistique X_i définie pour $z = (x_1, \ldots, x_n) \in \mathbb{R}^n$ par $X_i(z) = x_i$.

   S_1, S_2, S_3 sont des statistiques, avec $Z = (X_1, \ldots, X_n)$:
2. $S_1(Z) = \frac{X_1 + \ldots + X_n}{n}$, noté canoniquement $\bar{X}_n$
3. $S_2(Z) = \min\{X_i, i = 1 \ldots n\}$, et également les statistiques d'ordre $(X_{(i)})_{i=1,\ldots,n}$.
4. $S_3(Z) = \mathrm{med}(X_1, \ldots, X_n)$

Remarque : On adopte ici le point de vue probabiliste : les statistiques S_1, S_2, S_3 sont vues comme des variables aléatoires dérivées de $(X_1, \ldots, X_n)$. Alternativement, on peut les définir comme des fonctions $S_j : \mathbb{R}^n \to \mathbb{R}$, mesurables (par continuité) pour la tribu borélienne.

Définition : Statistiques indépendantes

Nous dirons que les statistiques S et T sur $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ sont indépendantes si pour toute loi $\mathbb{P} \in \mathcal{C}$, les éléments aléatoires S et T sont indépendants sous $\mathbb{P}$.

Exemple : Échantillon gaussien isotrope de moyenne $\mu \mathbf{1}_n$ (i.e. n-échantillon gaussien)

Considérons le modèle $(\mathsf{Z}, \mathcal{Z}, \mathcal{C}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_\theta\,,\ \theta \in \Theta := \mathbb{R} \times \mathbb{R}_+^*\})$ où pour tout $\theta \in \Theta$, $\mathbb{P}_\theta := p_\theta \cdot \mathrm{Leb}^{\otimes n}$ où p_θ est une densité gaussienne sur $\mathbb{R}^n$ de moyenne $\mu \mathbf{1}_n$ et de covariance $\sigma^2 \mathrm{I}_n$:

$$p_\theta(x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right)\,.$$

Notons X_i la i-ème observation. Sous $\mathbb{P}_\theta$, les statistiques $(X_1, \ldots, X_n)$ sont

- i.i.d. (indépendantes et identiquement distribuées)
- de densité gaussienne de moyenne $\mu \in \mathbb{R}$ et variance $\sigma^2 \in \mathbb{R}_+^*$:

$$p_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n q_\theta(x_i)\,, \quad q_\theta(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Loi image d'une statistique et modèle induit

Un "modèle induit" est le modèle obtenu en transformant nos observations.

Par exemple, si j'ai n v.a. $(X_1, \dots, X_n)$, je peux choisir de ne regarder que $\sum X_i$, ou X_1, etc.

Définition : Modèle induit

Soit T une statistique sur $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$.

- À une probabilité $\mathbb{P} \in \mathcal{C}$ on associe $\mathbb{P}^T$, la loi de T sous $\mathbb{P}$
- En écrivant $\mathcal{C}^T = \left\{\mathbb{P}^T\ ,\ \mathbb{P} \in \mathcal{C}\right\}$, on obtient ainsi le modèle statistique $(\mathsf{T}, \mathcal{T}, \mathcal{C}^T)$ induit par la statistique T.

Rappel : $\mathbb{P}^T$ est défini comme la loi image de $\mathbb{P}$ par T, définie, pour tout $A \in \mathcal{T}$, par

$$\mathbb{P}^T(A) = \mathbb{P}(T \in A)\ .$$

Calcul de la loi : méthode de la fonction muette : pour h mesurable positive sur T,

$$\int_{\mathsf{T}} h(t)\mathbb{P}^T(\mathrm{d}t) = \int_{\mathsf{Z}} h \circ T(z)\mathbb{P}(\mathrm{d}z) = \mathbb{E}[h(T)]\ .$$

Modèle induit : exemples

Exemple : Bernoulli

On considère

- le modèle $\left(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}\mathrm{er}(p)^{\otimes n}, p \in [0,1]\}\right)$
- la statistique $S_n = \sum_{i=1}^n X_i$.

Le modèle induit est :

$$\left(\{0,\dots,n\}, \mathcal{P}(\{0,\dots,n\}), \{\mathrm{Bin}(n,p), p \in [0,1]\}\right)$$

Exemple : Loi uniforme

On considère

- le modèle $\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathcal{U}([0;\theta])^{\otimes n}, \theta \in \mathbb{R}_+^*\}\right)$
- la statistique $X_{(n)} = \max X_i$.

(sous $\mathbb{P}_\theta$, chaque observation a pour densité par rapport à Lebesgue $p_\theta(t) = \frac{1}{\theta}\mathbb{1}_{[0;\theta]}(t)$)

Le modèle induit est :

$$\left(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_{n,\theta} \cdot \mathrm{d}\lambda, \theta \in \mathbb{R}_+^*\}\right)$$

avec $q_{n,\theta}(t) = \frac{nt^{n-1}}{\theta^n}\mathbb{1}_{[0;\theta]}(t)$ – **cf. PC de la semaine 4**

Définition : n-échantillon

Soient $(\mathsf{X}, \mathcal{X}, \mathcal{C})$ un modèle statistique et $n \in \mathbb{N}^*$. On appelle n-échantillon de $(\mathsf{X}, \mathcal{X}, \mathcal{C})$ le modèle statistique

$$(\mathsf{X}, \mathcal{X}, \mathcal{C})^n = (\mathsf{X}^n,\ \mathcal{X}^{\otimes n}, \{\mathbb{P} = \mathbb{Q}^{\otimes n},\ \mathbb{Q} \in \mathcal{C}\})\ .$$

On appelle i-ème observation canonique la statistique X_i

Rappel : notation $\mathbb{Q}^{\otimes n}$: mesure produit de $\mathbb{Q}$, n fois.

Proposition

$(X_1, \ldots, X_n) \sim \mathbb{Q}^{\otimes n}$ si et seulement si $(X_1, \ldots, X_n)$ sont i.i.d. de loi $\mathbb{Q}$. [2]

Démonstration

pour tout $\mathbb{P} = \mathbb{Q}^{\otimes n}$ avec $\mathbb{Q} \in \mathcal{C}$ et $(A_1, \ldots, A_n) \in \mathcal{X}^n$, nous avons

$$\mathbb{P}^Z(A_1 \times \cdots \times A_n) = \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{Q}^{\otimes n}(A_1 \times \cdots \times A_n) = \prod_{i=1}^n \mathbb{Q}(A_i) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i)\ .$$

Donc, en termes statistiques :

Lemme

Soit $(\mathsf{X}, \mathcal{X}, \mathcal{C})^n$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \mathcal{C})$.

1. Les observations $Z = (X_1, \ldots, X_n)$ sont indépendantes.
2. Pour tout $i \in \{1, \ldots, n\}$, le modèle induit par la statistique X_i est $(\mathsf{X}, \mathcal{X}, \mathcal{C})$.

Terminologie

"$(X_1, \ldots, X_n)$ ***est un n-échantillon du modèle*** $(\mathsf{X}, \mathcal{X}, \mathcal{C})$"

Approche et exemples

Le but du cours 2 sera d'explorer systématiquement les estimateurs possibles (en particulier par la méthode des moments et par le maximum de vraisemblance)
Néanmoins, dès cette semaine, on pourra "proposer" des estimateurs, typiquement basés sur des méthodes des moments simples.
Q1 : *Quelle est la probabilité qu'une pièce retombe sur "pile" ? Les pièces de "1ct" sont-elles biaisées ?* $\rightarrow$ estimation du paramètre d'une Bernoulli.
On considère un n-échantillon $X_1, \dots, X_n$ du modèle

$$(\{0,1\}, \mathcal{P}(\{0,1\}), \{\mathcal{B}\mathrm{er}(p), p \in [0,1]\}).$$

On souhaite estimer p.
Sous $\mathbb{P}_p$, on a $\mathbb{E}_p[X_1] = p$. On considère $\hat{p}_n = \bar{X}_n$.
Que peut-on dire de $\hat{p}_n$?

Figure – Wooclap

Exemple 2 : question sensible

Q3 : *Quelle proportion des 2A de Polytechnique a déjà utilisé un modèle d'IA pour faire une partie d'un DM alors que c'était interdit ?*

Je considère le même modèle que précédemment : $X_1, \dots, X_n$ du modèle

$$(\{0,1\}, \mathcal{P}(\{0,1\}), \{\mathcal{B}\mathrm{er}(p), p \in [0,1]\}).$$

Qu'en pensez-vous ?
⚠ les réponses vont être biaisées - on ne peut pas espérer observer $(X_1, \dots, X_n)$

Figure – Wooclap

On considère le protocole de *réponses randomisées* :

1. On pose une question sensible,
2. Chaque participant tire secrètement une variable aléatoire $U_i \sim \mathcal{B}\mathrm{er}(q)$, où q est connu (par exemple $q = 75\,\%$) ; les U_i sont i.i.d. et indépendantes des X_i,
3. si $U_i = 1$, il donne la vraie réponse, si $U_i = 0$, il répond à l'opposé.

En moyenne, les répondants donnent la vraie réponse avec probabilité q.

Estimateur par la méthode des moments

1. Sur le modèle des couples $(X_i, U_i)_{i=1,\dots,n}$, de loi $(\mathcal{B}\mathrm{er}(p) \otimes \mathcal{B}\mathrm{er}(q))^{\otimes n}$, on considère les statistiques $\mathrm{RR}_i = U_i X_i + (1-U_i)(1-X_i)$, $i = 1, \dots, n$.
2. Le modèle induit par $(\mathrm{RR}_1, \dots, \mathrm{RR}_n)$ est un n-échantillon du modèle $\big(\{0,1\},\ \mathcal{P}(\{0,1\}),\ \{\mathcal{B}(pq + (1-p)(1-q)),\ p \in [0,1]\}\big)$, q étant connu.

Sous le modèle induit $\mathbb{P}_p^{\mathrm{RR}}$,

$$\mathbb{E}_p[\mathrm{RR}_i] = (pq + (1-p)(1-q)) = p(2q-1) + 1 - q$$

donc, pour $q \neq 1/2$,

$$p = \frac{\mathbb{E}_p[\mathrm{RR}_i] - (1-q)}{2q-1}$$

On considère donc l'estimateur :

$$\hat{p}_n := \frac{\overline{\mathrm{RR}}_n - (1-q)}{2q-1}$$

nb : on peut même clipper sur $[0,1]$

Pour $q = 75\%$,

$$\hat{p}_n := 2(\overline{\mathrm{RR}}_n - 25\%).$$

On a $\mathbb{E}_p[\hat{p}_n] = p$ et, en notant $r = pq + (1-p)(1-q)$, $\mathrm{Var}_p(\hat{p}_n) = \frac{\mathrm{Var}_p(\overline{\mathrm{RR}}_n)}{(2q-1)^2} = \frac{r(1-r)}{n(2q-1)^2} \geq \frac{q(1-q)}{n(2q-1)^2}$.

Conclusion et programme de la PC

Objectifs du cours 1

- Comprendre le but du cours, le contexte, etc.
- Notion de modèle statistique
  - Motivation, Définition,
  - Identifiabilité, Modèle induit, modèle paramétrique, etc.
  - Notion de n-échantillon
  - Premiers exemples

⚠ Notions de probabilités clés, utilisées et à revoir : variable aléatoire, mesure produit, indépendance, densité, transfert et fonction muette !

Programme de la PC 1

Exercice	Notions du cours 1 & de Stats	Notions - rappels de probas
1 - *Un café pour commencer*	Modélisation, Estimation	Lois exponentielle et de Poisson
2 - *Transformation de v.a.*	Modèle induit	Fonction muette , Changement de variable
3 - *Modèle de translation et d'échelle*	Modèle induit, Th. de Gosset	Manipulation de lois, lois normales
4 - *Modèle de régression (linéaire) Gaussienne*	Th. de Cochran	Manipulation de lois, lois normales

nb : l'exercice 4 sera poursuivi en PC 2, le modèle de régression linéaire gaussienne étant un sujet qui reviendra régulièrement

Travail maison :

- 2 devoirs maison (une partie « problème », une partie numérique en Python).
- 3 à 4 Explorations Numériques - la première mise en ligne la semaine prochaine

→ Groupes de 3, tirés au sort.

Organisation et équipe enseignante

Cours en amphi : Aymeric Dieuleveut (séances 6 et 7 : Gersende Fort). **Petites classes :** l'équipe ci-dessous.

- **Aymeric Dieuleveut**, Professeur, École polytechnique, responsable du module
- **Gersende Fort**, Directrice de recherche CNRS, LAAS (Toulouse)
- **Sébastien Gadat**, Professeur, Toulouse School of Economics
- **Matthieu Lerasle**, Professeur, ENSAE
- **Angelina Roche**, Professeure, Université Paris Cité
- **Batiste Le Bars**, Chercheur Inria (équipe MAGNET, Lille)
- **Clément Bonet**, Professeur assistant, École polytechnique

Tutorat, explorations numériques et correction des DM :

- **Lucas Versini**, X21, doctorant au CMAP.

Figure – Le binet ENS recherche un délégué enseignement

Matériel

- Moodle : transparents (et version annotée après la séance), énoncés et corrigés de PC, DM, EN.
- Site du cours : cours, bases de maths, exercices, slides, animations.
- Polycopié 2026 : d'abord numérique (slide suivante).
- Tutorat : une séance d'une heure par semaine ; un sondage pour choisir l'horaire est sur Moodle.

À qui s'adresser ?

- DM, EN, tutorat : Lucas Versini.
- Organisation du cours : Aymeric Dieuleveut.
- Petites classes : votre chargé(e) de PC.


APM-41033-EP · INTRODUCTION AUX MÉTHODES STATISTIQUES · ÉCOLE POLYTECHNIQUE

Semaine 1 — Modèles statistiques

Observations, modèle statistique, identifiabilité, modèle induit, n-échantillon, premiers estimateurs.

Aymeric Dieuleveut — Professeur, École polytechnique

À propos · Syllabus · Semaine 1 · S2 · S3 · S4 · S5 · S6 · S7 · S8 · S9 · S10 · Devoirs · Équipe enseignante

Cours · Bases de maths · Exercices PC 1 · Slides PDF · Compléments (non exigibles)

L'objectif de ce chapitre est d'introduire la notion de **modèle statistique**. Concepts principaux : observations, modèle statistique, identifiabilité, statistique, modèle induit, modèle dominé, n-échantillon — et de premiers estimateurs élémentaires. Cet onglet reflète ce qui est couvert en amphi ; les prérequis sont dans l'onglet *Bases de maths*, la reste dans *Compléments*.

DANS CET ONGLET

Qu'est-ce qu'un modèle ?
1. Introduction : la démarche statistique
2. Modélisation statistique
3. Objectifs de la statistique
4. Identifiabilité
5. Statistiques
6. Loi image et modèle induit
7. Modèle statistique du n-échantillon
8. Exemples et premiers estimateurs
9. En pratique : programme des exercices

Qu'est-ce qu'un modèle ?

Moralement, un *modèle statistique* est une *description mathématique simplifiée d'un phénomène aléatoire réel*. Il sert de pont entre le monde des données, que l'on observe, et celui des lois de probabilité, qui représentent nos hypothèses sur la manière dont ces données sont produites. En choisissant un modèle, on fixe un cadre théorique dans lequel on pourra raisonner, estimer des grandeurs inconnues, prédire de nouvelles observations ou tester des hypothèses. Comme toute simplification, un modèle ne capture pas toute la réalité, mais en retient les aspects jugés essentiels pour répondre à une question précise.

Un modèle statistique est donc une *hypothèse mathématiquement formalisée* sur le mécanisme aléatoire ayant généré les données : un espace d'observations et une famille de lois de probabilité candidates, servant de cadre à l'inférence statistique.

1. Introduction : la démarche statistique

La statistique est l'étude de la *collecte*, de la *modélisation* et du *traitement* des données. Ses applications

Le polycopié devient d'abord numérique

- Le polycopié est réécrit cette année. Il vit d'abord sur le site du cours, semaine par semaine, aligné sur les slides : les fondamentaux, les bases de maths, les exercices, puis les compléments.
- La version PDF complète reste disponible. Une version papier peut être imprimée pour celles et ceux qui le souhaitent.
- Le site sera enrichi au fil des semaines : corrigés après chaque petite classe, animations, slides annotées.

But du cours - Introduction à la statistique mathématique

Spirit of the course

1. Introduction *avancée* à la statistique
2. Pour ceux qui poursuivront en maths appli, et pour les autres !
3. Notre approche - ne pas reculer devant un peu de formalisme pour gagner un peu de généralité
4. Aller en profondeur pour prouver de très beaux résultats

My *2 cents* :

1. Construire sur vos forces
2. Rien n'est difficile mais il y a du formalisme et du vocabulaire !
3. Ne prenez pas de retard !
4. Cours connu pour demander du travail - mais qui vous donnera d'excellentes bases !

Annexes

Slides supplémentaires : contenu

Ces slides supplémentaires détaillent quelques modèles intéressants.

1. Modèles de Survie
2. Modèle de régression
3. Systèmes de recommandation

5. Exemples détaillés de modèle et d'estimation
   - Modèles de Survie
   - Modèle de régression
   - Systèmes de recommandation

Modèle de survie

Modèle classique de durée de vie : famille de lois exponentielles de paramètre $\theta \in \mathbb{R}_+ \setminus \{0\}$, de densité par rapport à la mesure de Lebesgue Leb donnée par

$$q_\theta(x) = q(\theta, x) = \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) \ .$$

Le n-échantillon du modèle statistique

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{q_\theta \cdot \mathrm{Leb}\,,\ \theta \in \Theta = \mathbb{R}_+^*\})$$

est

$$\left(\mathbb{R}_+^n, \mathcal{B}(\mathbb{R}_+^n), \left\{\mathbb{P}_\theta = q_\theta^{\otimes n} \cdot \mathrm{Leb}^{\otimes n}\ ,\ \theta \in \Theta\right\}\right) \ ,$$

où $\mathrm{Leb}^{\otimes n}$ est la mesure de Lebesgue sur $\mathbb{R}^n$ et

$$q_\theta^{\otimes n}(x_1, \ldots, x_n) = \prod_{i=1}^n q(\theta, x_i).$$

Absence de mémoire : pour tout $a, b > 0$ et $\theta \in \mathbb{R}_+^*$,

$$\mathbb{P}_\theta\left(X_1 \geq a + b \,|\, X_1 \geq a\right) = \frac{\mathrm{e}^{-(a+b)\theta}}{\mathrm{e}^{-a\theta}} = \mathbb{P}_\theta(X_1 \geq b)$$

L'observation $(X_1, \ldots, X_n)$ est définie par $X_i(x_1, \ldots, x_n) = x_i$ pour tout $i \in \{1, \ldots, n\}$ et $(x_1, \ldots, x_n) \in \mathbb{R}_+^n$.

- Le modèle statistique induit par X_i est

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{q_\theta \cdot \mathrm{Leb}\,,\ \theta \in \Theta = \mathbb{R}_+^*\})\ .$$

- Les statistiques X_i et X_j sont indépendantes pour $i \neq j$.

Calcul des moments - loi exponentielle

- fonction génératrice des moments : pour $|s| < \theta$,

$$\begin{aligned}\mathbb{E}_\theta[\mathrm{e}^{sX}] &= \int_0^\infty \mathrm{e}^{sx}\theta\mathrm{e}^{-\theta x}\mathrm{d}x \\ &= \frac{\theta}{\theta - s} = \sum_{k=0}^{\infty} \frac{s^k}{\theta^k}\end{aligned}$$

- Développement de l'exponentielle $\mathbb{E}_\theta[\mathrm{e}^{sX}] = \sum_{k=0}^{\infty} \frac{s^k}{k!}\mathbb{E}_\theta[X^k]$
- En identifiant les développements, on obtient directement :

$$\mathbb{E}_\theta[X^k] = \frac{k!}{\theta^k}\,, \quad k \in \mathbb{N}^*\,.$$

Mise en pratique : estimateurs des moments – loi exponentielle

Considérons les fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$.

$$e(\theta) = \int T(x)\mathbb{Q}_\theta(\mathrm{d}x) = \int_0^{+\infty} x\theta \exp(-\theta x)\mathrm{d}x = \frac{1}{\theta}$$
$$\tilde{e}(\theta) = \int \tilde{T}(x)\mathbb{Q}_\theta(\mathrm{d}x) = \int_0^{+\infty} x^2\theta \exp(-\theta x)\mathrm{d}x = \frac{2}{\theta^2}\ .$$

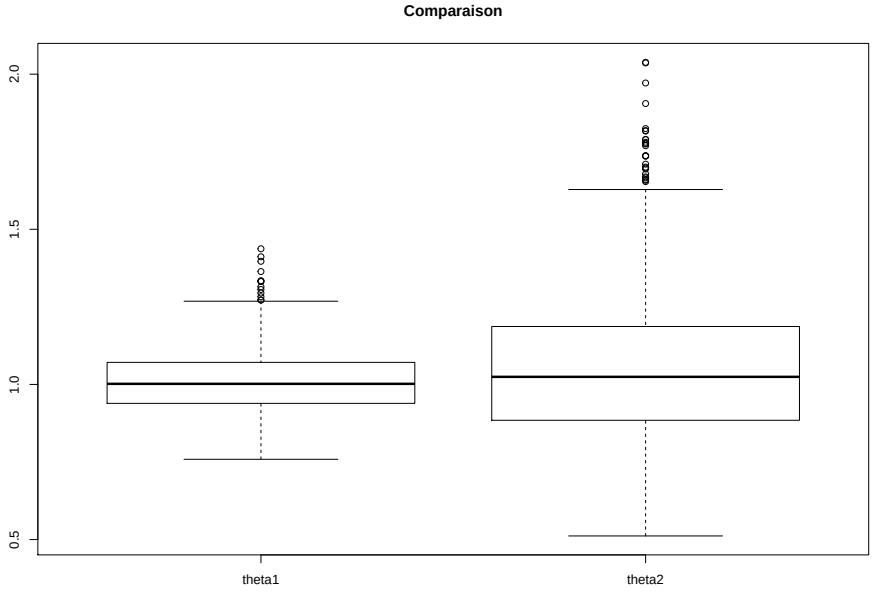
- Les estimateurs des moments associés sont les solutions des équations

$$e(\theta) = \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{et} \quad \tilde{e}(\theta) = \frac{2}{\theta^2} = \frac{1}{n}\sum_{i=1}^n X_i^2\ .$$

  Bien entendu, chacun des moments utilisés donnera dans ce contexte un estimateur différent.

- Ces équations ont des solutions uniques :

$$\hat{\theta}_{n,1} = \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i} \quad \text{et} \quad \hat{\theta}_{n,2} = \left(\frac{2}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)^{1/2}\ .$$


Comparaison
2.0
1.5
1.0
0.5
theta1
theta2

Définition : Lois Gamma

- Pour tout réel $p > 0$, on appelle *loi Gamma réduite* de paramètre de forme p (et l'on note $\mathrm{Gamma}(p)$) la loi définie sur l'ensemble des réels positifs par la densité

$$f_p(x) = \frac{1}{\Gamma(p)} \exp(-x)\ x^{p-1}, \quad x > 0\ .$$

- Pour $\lambda > 0$, on appelle $\mathrm{Gamma}(p, \lambda)$, la loi de $X = Z/\lambda$ où $Z \sim \mathrm{Gamma}(p)$ et λ est le paramètre d'intensité (l'échelle est $1/\lambda$). La densité de la loi $\mathrm{Gamma}(p, \lambda)$ est donnée par

$$f_{p,\lambda}(x) = \frac{\lambda^p}{\Gamma(p)} \exp(-\lambda x)\ x^{p-1}\ , \quad x > 0\ .$$

La loi exponentielle d'intensité $\lambda > 0$ est la loi $\mathrm{Gamma}(1, \lambda)$.
Si Z suit une loi $\mathrm{Gamma}(p)$, nous avons pour tout $r > -p$,

$$\mathbb{E}[Z^r] = \frac{\Gamma(p+r)}{\Gamma(p)}\ .$$

Lemme : Somme de variables Gamma indépendantes

Soit $(X_1, \ldots, X_n)$ n variables aléatoires indépendantes distribuées suivant des lois $\mathrm{Gamma}(p_i, \theta)$ avec $\theta > 0$ et $p_i > 0$, $i \in \{1, \ldots, n\}$. Alors, $\sum_{i=1}^{n} X_i$ est distribuée suivant une loi $\mathrm{Gamma}(\sum_{i=1}^{n} p_i, \theta)$.

En particulier, la somme $X_1 + X_2$ de deux variables aléatoires indépendantes de même loi exponentielle $\mathcal{E}(\theta)$ suit une loi $\mathrm{Gamma}(2, \theta)$.

$$\mathcal{E}(\theta) \star \mathcal{E}(\theta) = \mathrm{Gamma}(2, \theta)$$

Propriétés de l'estimateur des moments (loi exponentielle) $n/\sum_{i=1}^n X_i$

- On sait que sous $q_\theta \cdot \mathrm{Leb}$, $\sum_{i=1}^n X_i$ suit une loi $\mathrm{Gamma}(n,\theta)$. Donc, pour $n \geq 3$ (l'espérance existe dès $n \geq 2$),

$$\mathbb{E}_\theta[\hat{\theta}_{n,1}] = \mathbb{E}_\theta\left[n/\sum_{i=1}^n X_i\right] = \int_0^\infty \frac{\theta^n}{(n-1)!}\mathrm{e}^{-\theta y} y^{n-1}\frac{n}{y}\mathrm{d}y = \frac{n}{n-1}\theta$$

$$\begin{aligned}\mathbb{E}_\theta[(\hat{\theta}_{n,1}-\theta)^2] &= \mathbb{E}_\theta[\hat{\theta}_{n,1}^2] - 2\theta\mathbb{E}_\theta[\hat{\theta}_{n,1}] + \theta^2 \\ &= \theta^2\frac{n^2}{(n-1)(n-2)} - 2\theta^2\frac{n}{n-1} + \theta^2 = \frac{n+2}{n-1}\frac{\theta^2}{n-2}\end{aligned}$$

- Il est important de remarquer que les calculs précédents sont menés directement en utilisant des renormalisations de loi Gamma.
- On obtient comme carré de l'erreur en moyenne

$$\mathbb{E}_\theta[(\hat{\theta}_{n,1}-\theta)^2] = \frac{n+2}{(n-1)(n-2)}\theta^2$$

  Cette erreur est en moyenne d'autant plus grande que θ est grand.
- On pourrait mener les mêmes calculs pour le second estimateur des moments $\hat{\theta}_{n,2}$.

Anomalie température moyenne

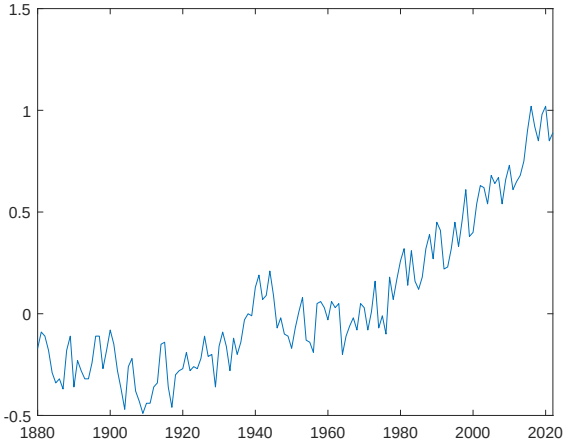


Figure – anomalie de température moyenne à l'échelle du globe par rapport à la période 1951-1980, NASA, Goddard Institute for Space Studies

Construction du modèle

- Observations : déviation des températures annuelles par rapport à la moyenne 1951-1980 : $(x_1, \ldots, x_n) \in \mathbb{R}^n$
- Modèle statistique paramétrique

$$p_{\boldsymbol{\beta},\sigma^2}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left\{x_i - m(\boldsymbol{\beta}, i)\right\}^2\right)$$

où $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ est un paramètre et pour tout $\boldsymbol{\beta} \in \mathbb{R}^p$, $i \mapsto m(\boldsymbol{\beta}, i)$ est une fonction.

- Paramètres $\theta = (\beta_0, \ldots, \beta_{p-1}, \sigma^2) \in \Theta = \mathbb{R}^p \times \mathbb{R}_+^*$

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{p_\theta \cdot \mathrm{Leb}^{\otimes n}, \theta \in \Theta\})$$

Hypothèses du modèle

Ici $\mathsf{Z} = \mathbb{R}^n$ et $Z = (X_1, \ldots, X_n)$ où $X_i(x_1, \ldots, x_n) = x_i$ pour $i \in \{1, \ldots, n\}$

- Pour tout $\theta \in \Theta$, X_i et X_j, $i \neq j$, sont indépendants sous $\mathbb{P}_\theta$
- Sous $\mathbb{P}_\theta$ et tout $i \in \{1, \ldots, n\}$, la distribution de X_i est gaussienne de variance σ^2
- Sous $\mathbb{P}_\theta$ et tout $i \in \{1, \ldots, n\}$, la moyenne de X_i est $m(\boldsymbol{\beta}, i)$

Ce sont des hypothèses de modélisation, qui peuvent être discutées etc.

- Est-il raisonnable de supposer que la variance est constante (homoscédasticité) ?
- A-t-on une bonne raison de penser que la distribution des erreurs est Gaussienne ?
- Les observations sont-elles indépendantes ?

Tendance linéaire

- Hypothèse 1 : tendance linéaire sur l'ensemble de la période :

$$m(\boldsymbol{\beta}, i) = \beta_0 + \beta_1 t(i)$$

où $t(i)$ est l'index de temps

- Construction des estimateurs : on estime les paramètres en minimisant l'erreur quadratique

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^2} \sum_{i=1}^{n} (X_i - \beta_0 - \beta_1 t(i))^2$$

$$\hat{\sigma}^2 = (n-2)^{-1} \sum_{i=1}^{n} (X_i - \hat{\beta}_0 - \hat{\beta}_1 t(i))^2 \quad \text{(version sans biais)}$$

$\hat{\boldsymbol{\beta}}$ est aussi, dans le modèle gaussien, l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$; l'EMV de σ^2 est $n^{-1} \sum_{i=1}^{n} (X_i - \hat{\beta}_0 - \hat{\beta}_1 t(i))^2$, dont $\hat{\sigma}^2$ est la version sans biais :

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \max_{\sigma^2 > 0} p_{\boldsymbol{\beta}, \sigma^2}(X_1, \ldots, X_n)$$

Solution

La solution est explicite !

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (t(i) - \bar{t})(X_i - \bar{X}_n)}{\sum_{i=1}^n (t(i) - \bar{t})^2}\,, \qquad \widehat{\beta}_0 = \bar{X}_n - \widehat{\beta}_1\,\bar{t}\,, \qquad \bar{t} = \frac{1}{n}\sum_{i=1}^n t(i)$$

(il faut au moins deux dates distinctes, i.e. Φ de rang 2)
ou, de façon plus compacte,

$$\widehat{\boldsymbol{\beta}} = (\Phi^\top \Phi)^{-1}\Phi^\top \mathbf{X} \quad \text{avec } \Phi = \begin{bmatrix} 1 & t(1) \\ \vdots & \vdots \\ 1 & t(n) \end{bmatrix} \text{ et } \mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

La variance est donnée par

$$\hat{\sigma}^2 = (n-2)^{-1}\sum_{i=1}^n \{X_i - \widehat{\beta}_0 - \widehat{\beta}_1 t(i)\}^2 = (n-2)^{-1}\|\mathbf{X} - \Phi\hat{\boldsymbol{\beta}}\|^2$$

La distribution de ces estimateurs est connue : pour tout $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+^*$, sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\Phi^\top\Phi)^{-1})$$

$$(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)\,, \quad \widehat{\boldsymbol{\beta}} \text{ et } \hat{\sigma}^2 \text{ indépendants (théorème de Cochran)}$$

Tendance linéaire

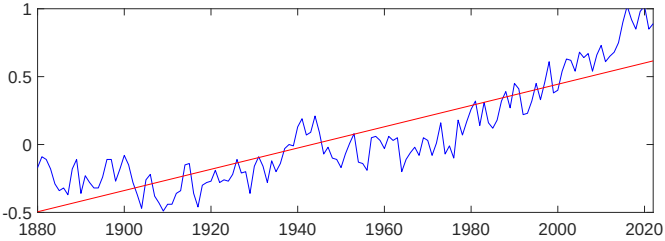
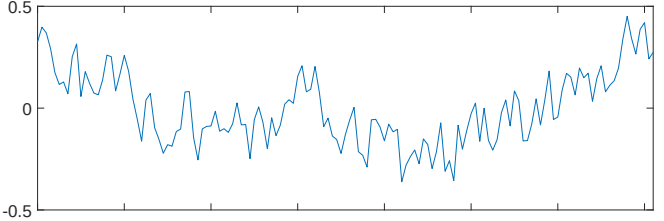


Figure – anomalie de température moyenne à l'échelle du globe par rapport à la période 1951-1980, NASA, GISS. Estimation et écart-type estimé : $\widehat{\beta}_1 \pm \mathrm{se}_{\beta_1} = 0{,}0072\,^\circ\mathrm{C}$ par an $\pm\ 0{,}0008\,^\circ\mathrm{C}$ par an (intervalle de confiance à 95 % : $\widehat{\beta}_1 \pm 1{,}96\,\mathrm{se}_{\beta_1}$)



Modèle avec rupture de pente

- Modèle 2 : rupture de pente

$$m(\boldsymbol{\beta}, i) = \beta_0 + \beta_1 t(i) + \beta_2 \left(t(i) - t(\beta_3)\right) \mathbb{1}_{\{i > \beta_3\}}, \qquad \beta_3 \in \{1, \dots, n\}$$

- Construction des estimateurs Minimisation de l'erreur quadratique

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^3 \times \{1,\dots,n\}}{\arg\min} \sum_{i=1}^{n} (X_i - m(\boldsymbol{\beta}, i))^2$$

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - m(\hat{\boldsymbol{\beta}}, i))^2$$

qui est aussi l'estimateur du maximum de vraisemblance

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta}, \sigma^2}{\arg\max}\, p_{\boldsymbol{\beta}, \sigma^2}(X_1, \dots, X_n)$$

Pour β_3 - position de la rupture de pente - fixé, les paramètres $(\beta_0, \beta_1, \beta_2)$ peuvent être calculés de façon explicite

$$\hat{\boldsymbol{\beta}}(\beta_3) = \begin{bmatrix} \hat{\beta}_0(\beta_3) \\ \hat{\beta}_1(\beta_3) \\ \hat{\beta}_2(\beta_3) \end{bmatrix} = (\Phi(\beta_3)^\top \Phi(\beta_3))^{-1} \Phi(\beta_3)^\top \mathbf{X}$$

où

$$\Phi(\beta_3) = \begin{bmatrix} 1 & t(1) & 0 \\ 1 & \vdots & \vdots \\ 1 & t(\beta_3) & 0 \\ 1 & t(\beta_3+1) & t(\beta_3+1) - t(\beta_3) \\ \vdots & \vdots & \vdots \\ 1 & t(n) & t(n) - t(\beta_3) \end{bmatrix}$$

On estime β_3 en maximisant la vraisemblance profilée, c'est-à-dire en minimisant la somme des carrés des résidus :

$$\hat{\beta}_3 = \underset{\beta_3 \in \{1,\ldots,n\}}{\arg\min} \|\mathbf{X} - \Phi(\beta_3)\hat{\boldsymbol{\beta}}(\beta_3)\|^2$$

Modèles avec rupture de pente

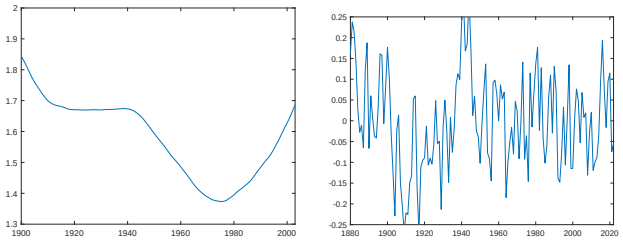


Figure – Gauche : vraisemblance profilée - Droite : résidu de prédiction

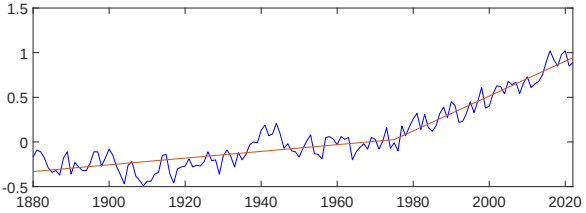


Figure – anomalie de température moyenne à l'échelle du globe par rapport à la période 1951-1980, NASA, GISS

					
Claire	4	?	3	...	
Nial	?	4	?	...	
Brendon	?	5	4	...	
Andrew	?	4	?	...	
Adrian	1	?	?	...	
Damien	?	1	?	...	
⋮	⋮	⋮	⋮	⋱	

Une formalisation statistique possible

- m_1 : le nombre de sujets
- m_2 : le nombre d'objets (films, produits, etc..)
- Y : matrice $m_1 \times m_2$ avec

$$Y_{i,j} = \text{ note du sujet } i \text{ au produit } j$$

- données : $Y_{i,j}$ pour $(i,j) \in I$

$$n = \text{ nombre d'observations } = \mathrm{card}(I) \ll m_1 m_2$$

Une formalisation statistique possible

- Les notes $Y_{i,j}$ sont indépendantes, à valeurs dans $\{1, \dots, K\}$, et distribuées suivant la loi (catégorielle)

$$\frac{\exp(\theta_{i,j,1})}{\sum_{k=1}^{K} \exp(\theta_{i,j,k})}, \dots, \frac{\exp(\theta_{i,j,K})}{\sum_{k=1}^{K} \exp(\theta_{i,j,k})}$$

  Par convention, on fixe $\theta_{i,j,K} = 0$ autrement le modèle n'est pas identifiable

- La densité de l'observation par rapport à la mesure de comptage est

$$p_\theta(\{y_{i,j}\}_{(i,j)\in I}) = \prod_{(i,j)\in I} \prod_{\ell=1}^{K} \left(\frac{\exp(\theta_{i,j,\ell})}{\sum_{k=1}^{K} \exp(\theta_{i,j,k})} \right)^{\mathbb{1}_{\{y_{i,j}=\ell\}}} .$$

Une hypothèse sur la structure de θ

- Nécessité de faire une hypothèse sur la structure de θ, qui soit raisonnable pour l'application considérée
- plusieurs possibilités ! Par exemple, si les notes sont binaires ($K = 2$, un seul paramètre $\theta_{i,j} = \theta_{i,j,1}$ par couple), les préférences des utilisateurs sont des mélanges d'un petit nombre de comportements « prototypes », i.e. la matrice $\theta = (\theta_{i,j})$ est de rang $\leq q$:

$$\theta = UV^\top$$

où

  - U est une matrice $m_1 \times q$,
  - V est une matrice $m_2 \times q$.

- Le nombre total de paramètres est égal à $q(m_1 + m_2)$ et le rang q est souvent choisi de telle sorte que

$$q(m_1 + m_2) \ll n \ll m_1 m_2$$

Recommandation

- Estimer les matrices U et V... Dans ce cas (comme dans la plupart des applications "réelles"), il n'y a pas de méthodes d'estimation "élémentaires" (voir Cours 2 méthodes d'estimation)
- Imputer les données manquantes.
- C'est la base des systèmes de recommandations ou de filtrage collaboratif (dans un système "opérationnel", il y a bien entendu un certain nombre de raffinements à apporter, mais c'est la "base").