

# MDI220, Statistique

## Cours 1

Equipe pédagogique: Thomas Bonald , Anas Barakat, Anne Sabourin,  
Umut Simsekli, Guillaume Staerman.

Septembre 2019

# Chapitre 1

## Introduction : analyse statistique de données

### 1. Exemples

Proportion de défauts

Test A/B

### 2. Formalisation

Cadre probabiliste

Modèle statistique, paramétrisation

### 3. Théorie de la décision : concepts de base

Actions, fonction de coût

Risque

Exemples de coûts et risques associés.

## Faire des statistiques ?

- Utiliser les données pour **apprendre** / **extraire de l'information** sur leur distribution probabiliste
- Beaucoup de données → beaucoup d'applications  
*... Big data ... .. Data Science ... .. Machine Learning ...  
Intelligence artificielle ...*
- **ce cours** : vous donner outils théoriques + mise en oeuvre pratique pour attaquer problèmes de
  - procédures d'estimation
  - tests, intervalles de confiance
  - ...→ procédures d'« inférence » ou « apprentissage » sur lesquels reposent beaucoup d'algorithmes.

# Fonctionnement du cours

- Mardi matin
- TH1 : cours en amphi ; TH2 : TD (présence obligatoire ? cf votre prof de TD)  
(Aujourd'hui : 2eme TH = TP, prise en main du logiciel R)
- poly : en ligne, version papier : imminente.
- TDs : en ligne au fil du cours + corrigés.
  
- Evaluation :
  - Contrôle continu + mini projet : 40%
  - Contrôle final : 60% → droit à 1 feuille manuscrite de notes, pas de poly ni TD
  
- Mini-projet :
  - individuel
  - $\approx$  un gros DM avec théorie + code (R, Rstudio, cf. TP aujourd'hui)
  - rendu : un notebook + scripts

## 1. Exemples

Proportion de défauts

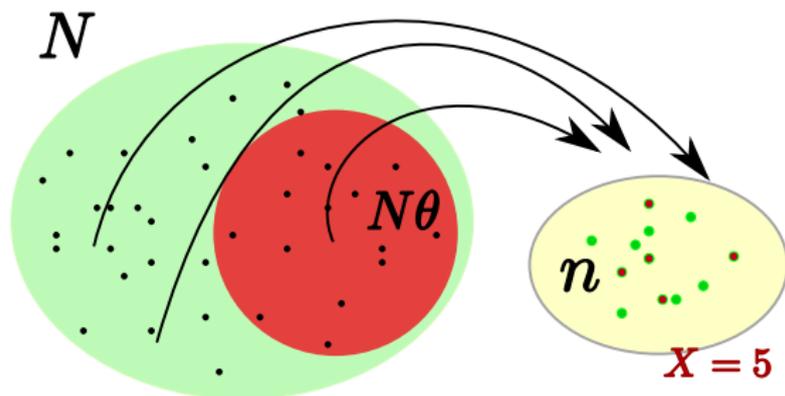
Test A/B

## 2. Formalisation

## 3. Théorie de la décision : concepts de base

## Proportion de défauts dans un population

- $N$  individus/clients/pièces produites par une machine
- proportion  $\theta$  de défectueux,  $\theta \in \{1/N, \dots, (N-1)/N, 1\}$ .
- **expérience** : tirage aléatoire, uniforme, sans remise, de  $n$  individus parmi  $N$
- **on observe** :  $X$  : nombre de défauts parmi les  $n$  tirés.



Comment utiliser  $X$  pour apprendre qqchose de  $\theta$  ?

## Objectifs possibles

- Tester si  $\theta < 5\%$
- Estimer  $\theta$  (construire un estimateur  $\hat{\theta}$ )
- Donner un intervalle de confiance contenant  $\theta$  avec grande probabilité.

# Modélisation

à  $\theta$  fixé, comment se comporte  $X$  ?

**Idee de la suite** : résoudre ensuite un « problème inverse » pour retrouver  $\theta$  à partir de  $X$

- $\mathbb{P}_\theta$  : proba sous jacente lorsque le paramètre vaut  $\theta$ .
- Calcul de  $\mathbb{P}_\theta(X = k)$  :

**au tableau**

- $\mathbb{P}_\theta(X = k) =$  une fonction de  $k, \theta, n, N$

On a défini un « modèle statistique », *i.e.* une famille de lois de probas possibles pour  $X$  (une pour chaque valeur de  $\theta$ ).

- On supposera que  $X \sim \mathbb{P}_\theta$ , avec  $\theta$  inconnu.

## Estimation de $\theta$

- Ayant observé «  $X = x$  », peut-on donner une estimation  $\hat{\theta}$  de  $\theta$  ?
- **idée 1** : calculer l'espérance théorique de  $X$  à  $\theta$  fixé,  $\mathbb{E}_\theta(X)$ , et ajuster  $\hat{\theta}$  pour avoir  $x = \mathbb{E}_{\hat{\theta}}(X)$ .
- ... gros calcul ...

$$\mathbb{E}_\theta(X) = n\theta, \quad \forall \theta \in \{0, 1/N, 2/N, \dots, 1\}.$$

→ estimateur « naturel » de  $\theta$  :

$$\hat{\theta} = \frac{X}{n}$$

**N.B.** :  $\hat{\theta}$  est une fonction de  $X$

## Estimation de $\theta$ (ii)

- **idée 2** : puisque  $X \leq N\theta$ , prendre

$$\hat{\theta}_2 = \frac{X}{N}$$

(ainsi : on est sûrs de ne pas sur-estimer  $\theta$ )

- **idée 3** : : puisque  $n - X \leq N - N\theta$ , prendre

$$\hat{\theta}_3 = \frac{N - (n - X)}{N}$$

(ainsi : on est sûrs de ne pas sous-estimer  $\theta$ )

- **idée 4** :  $\hat{\theta}_3 = \frac{1}{2} \left[ \frac{X}{N} + \frac{N - (n - X)}{N} \right]$
- tous ces estimateurs sont des **fonctions de  $X$**
- choix : en fonction du “risque” attaché à chaque estimateur (dépend de la préférence de l'utilisateur).

## 1. Exemples

Proportion de défauts

Test A/B

## 2. Formalisation

## 3. Théorie de la décision : concepts de base

## Effacité d'un traitement/ une stratégie marketing/ ...

- Le traitement/la stratégie est-il efficace?
- 2 échantillons  $(X_1, \dots, X_n) \stackrel{i.i.d.}{\sim} F, (Y_1, \dots, Y_n) \stackrel{i.i.d.}{\sim} G$
- Question :  $F = G$ ?
- besoin de faire des hypothèses, ex :
  - $Y_i \stackrel{\text{loi}}{=} X_i + \Delta$ , i.e.  $G(\cdot) = F(\cdot - \Delta)$
  - $Y_i \sim \mathcal{N}(\mu + \Delta, \sigma^2), X_i \sim \mathcal{N}(\mu, \sigma^2), \mu$  inconnu,  $\sigma^2$  connu
  - ...
- choix de modèle : problème récurrent !

1. Exemples

2. Formalisation

Cadre probabiliste

Modèle statistique, paramétrisation

3. Théorie de la décision : concepts de base

# Notations (I)

- Univers  $\Omega$ , réalisation  $\omega \in \Omega$ .
- tribu  $\mathcal{F}$  sur  $\Omega$  : ensemble des événements  
(un événement = un élément de  $\mathcal{F}$  = un sous ensemble de  $\Omega$ )
- Espace des observations :  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$   
tribu  $\mathcal{B}(\mathcal{X})$  :  $\mathcal{P}(\mathcal{X})$  si  $\mathcal{X}$  discret, boréliens si  $\mathcal{X} = \mathbb{R}^d$ .
- Observation : variable aléatoire (*i.e.* une fonction mesurable)  
 $X : \Omega \rightarrow \mathcal{X}$ .

## Notations(II)

- loi de  $X$  :  $P$ . C'est une probabilité sur  $\mathcal{X}$  :

$$\begin{aligned}\forall A \in \mathcal{B}(\mathcal{X}), \quad P(A) &= \mathbb{P}(X \in A) \\ &= \mathbb{P}\{\omega : X(\omega) \in A\} \\ &= \mathbb{P} \circ X^{-1}(A) \\ &= \text{« mesure image de } P \text{ par } X \text{ »}\end{aligned}$$

- on écrit  $X \sim P$ .
- En statistique, la proba sous-jacente  $\mathbb{P}$  est inconnue, donc  $P$  aussi.
- **But** : obtenir de l'info sur  $P$  en observant  $X$ .

1. Exemples

2. Formalisation

Cadre probabiliste

Modèle statistique, paramétrisation

3. Théorie de la décision : concepts de base

## modèle statistique

modèle = connaissance « a priori » du statisticien (avant l'expérience)  
→ famille de lois « possibles » pour  $X$  :

### **definition : modèle statistique**

Un modèle statistique est une famille de lois de probabilités, notée  $\mathcal{P}$  :

$$\mathcal{P} \subset \{ \text{toutes les lois de proba sur } \mathcal{X} \}$$

Lors de l'expérience statistique, on suppose que  $X \sim P$  avec  $P \in \mathcal{P}$ .

## paramétrisation, espace des paramètres

### définition : espace des paramètres, paramétrisation

paramétrisation : application

$$\Theta \rightarrow \mathcal{P}$$

$$\theta \mapsto P_\theta$$

où  $\Theta$  est un ensemble appelé "espace des paramètres".

(paramétrisation = **étiquetage** des lois  $P \in \mathcal{P}$  : par un **paramètre**  $\theta \in \Theta$ , supposé facile à manipuler (ex :  $\theta \in \mathbb{R}^d$ ))

ex : espace des paramètres pour le modèle contenant toutes les lois normales ?

## paramétrique/ non paramétrique

- modèle **paramétrique** :  $\exists$  une paramétrisation telle que  $\Theta \subset \mathbb{R}^d$
  - exemple ?
- 
- modèle **paramétrique** :  $\exists$  de paramétrisation telle que  $\Theta \subset \mathbb{R}^d$
  - exemple : ensemble de toutes les lois de probas à densité, dont la densité est « symétrique » (une fonction paire).

## Notations : variable aléatoire dans un modèle

- on note

$$X \sim P_\theta, \theta \in \Theta.$$

( $X$  suit la loi  $P_\theta$ ), où (généralement)  $\theta$  est fixé mais n'est pas observé et  $X$  est observé.

- **N.B.** même en non paramétrique, on peut toujours choisir  $\Theta = \mathcal{P}$  et écrire

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}.$$

→ pas de problèmes de notations

## Travail du statisticien

- La seule connaissance mise à la disposition du statisticien est un modèle  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$  et une réalisation de l'observation  $X \sim P_\theta$ , où  $\theta \in \Theta$  est inconnu.
- L'objectif est d'approcher une certaine quantité d'intérêt  $g(\theta)$  (dépendant uniquement de  $\theta$ ) en utilisant une procédure fondée uniquement sur l'observation  $X$  (une fonction ne dépendant que de  $X$ ).

## Quantités d'intérêt $g(\theta)$ usuelles

- intervalle contenant  $\theta$
- $\mathbb{P}_\theta(X > u)$  ( $u$  : seuil à risque)
- $\mathbb{E}_\theta(X)$
- $\mathbb{1}_{\Theta_0}(\theta)$  où  $\Theta_0 \subset \Theta$ .

souvent :  $g(\theta)$  est aussi appelée 'paramètre' (d'intérêt) même si  $g(\theta)$  ne détermine pas entièrement  $\mathbb{P}_\theta$ .

## Notion de 'statistique'

Toute l'inférence doit se faire à partir des données seulement :

### définition : statistique

Une *statistique* est une variable aléatoire s'écrivant comme une fonction mesurable des observations, de type  $\varphi(X)$  où

$\varphi : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \rightarrow (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$  est une fonction mesurable.

en particulier

### definition : estimateur

Un *estimateur* d'une quantité  $g(\theta) \in \mathcal{G} \subset \mathbb{R}^d$  est une statistique

$\hat{g} : x \in \mathcal{X} \mapsto \hat{g}(x) \in \mathcal{G}$ .

# Exemples

- nombre de défauts :
  - modèle ?  $\Theta$  ?
  - modèle paramétrique ou non ?
- modèle 'semi-paramétrique' :

$$\mathcal{P} = \{ \text{lois de densité } f(\cdot - \mu), \text{ où} \\ f : \text{densité paire sur } \mathbb{R} \\ \mu \in \mathbb{R} \}$$

$$\Theta = \{ (f, \mu) : f \text{ densité paire}, \mu \in \mathbb{R} \}$$

Quand a-t-on une chance de “retrouver”  $\theta$  à partir des observations ?

## définition : identifiabilité.

- La paramétrisation  $\theta \mapsto P_\theta$  est dite identifiable si elle est injective. (i.e.  $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$ ).
- une grandeur d'intérêt  $g(\theta)$  est dite identifiable si

$$g(\theta_1) \neq g(\theta_2) \quad \Rightarrow \quad P_{\theta_1} \neq P_{\theta_2}.$$

## Modèle dominé

### définition : modèle dominé

le modèle  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$  est appelé **dominé**, si toutes les lois  $P_\theta, \theta \in \Theta$  admettent une densité par rapport à une **même** mesure de référence  $\sigma$ -finie\*  $\mu$ ,

\*  $\sigma$ -finie : l'espace  $\mathcal{X}$  est une union dénombrable d'ensembles de mesure finie.

- **Rappel** Une probabilité  $P$  sur  $\mathcal{X}$  admet une densité  $f$  par rapport à une mesure  $\mu$  sur  $\mathcal{X}$ , si

$$\forall A \in \mathcal{B}(\mathcal{X}), P(A) = \int_A f(x) d\mu(x)$$

## Rappels : densité/ mesure

- Caractérisations équivalentes d'une densité :

$$(i) \quad \forall A \in \mathcal{B}(\mathcal{X}), P(A) = \int_A f(x) d\mu(x)$$

$$(ii) \quad \forall \phi : \mathcal{X} \rightarrow \mathbb{R}^+ \text{ (mesurable), } \mathbb{E}(\phi) = \int_{\mathcal{X}} \phi(x) f(x) d\mu(x)$$

$$(iii) \text{ (Radon-Nikodym)} \quad \forall A \in \mathcal{B}(\mathcal{X}), \mu(A) = 0 \Rightarrow P(A) = 0.$$

### exemples

- $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$  :  $\mu$  = mesure de Lebesgue sur  $\mathbb{R}$ .
- $\mathcal{P} = \{\text{Poisson}(\lambda), \lambda > 0\}$  :  $\mu$  = mesure de comptage sur  $\mathbb{N}$ .

### contre-exemple

- $\mathcal{P} = \{\delta_x, x \in \mathbb{R}\}$  : seule  $\mu$  possible : comptage sur  $\mathbb{R}$ ,  $\mu(A) = |A|$  (infini dès que  $A$  contient un intervalle) : pas  $\sigma$ -fini !

- Ici trois cas possibles (suffisent pour comprendre)

1.  $\mu =$  Lebesgue, sur  $\mathbb{R}$  : alors

$$\int_{\mathbb{R}^d} \Phi(x) d\mu(x) = \int_{\mathbb{R}^d} \Phi(x) dx \quad (\text{intégrale de Riemann, si elle existe.})$$

2.  $\mu$  : mesure de comptage sur  $\mathcal{X}$  discret,  $\mu = \sum_{x \in \mathcal{X}} \delta_x$  (Diracs)

$$\int_{\mathcal{X}} \Phi(x) d\mu(x) = \sum_{x \in \mathcal{X}} \Phi(x).$$

3. Mélange des deux :  $\mathcal{X} \subset \mathbb{R}^d$ ,  $\mu = \mu_1 + \mu_2$  où :  
 $\mu_1$  : comptage sur  $\mathcal{X}_0 \subset \mathcal{X}$  discret et  $\mu_2$  : Lebesgue sur  $\mathbb{R}^d$ .

$$\int_{\mathcal{X}} \Phi(x) d\mu(x) = \sum_{x \in \mathcal{X}_0} \Phi(x) + \int_{\mathbb{R}^d} \Phi(x) dx.$$

# Vraisemblance

- Dans un modèle dominé  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ .
- chaque  $P_\theta$  admet une densité  $p_\theta : \mathcal{X} \rightarrow \mathbb{R}^+$  par rapport à la mesure de référence  $\mu$ .

## définition : vraisemblance

Soit  $x \in \mathcal{X}$  une observation. L'application

$$\theta \mapsto p_\theta(x)$$

est appelée “fonction de vraisemblance”, ou “vraisemblance” associée à  $x$ .

- Attention : la vraisemblance est une fonction de  $\theta$ , à  $x$  fixé.
- Intérêt : d'habitude, (en probas) on s'intéresse à  $p_\theta(\cdot)$ , à  $\theta$  fixé.
- en stats : on ne “voit” que  $x$ , on cherche  $\theta$ .
- à  $x$  fixé,  $\theta$  est d'autant plus “vraisemblable” que  $p_\theta(x)$  est élevé.
- un moyen d'estimer  $\theta$  est de maximiser en  $\theta$  (à  $x$  fixé), la vraisemblance. cf. estimateur du maximum de vraisemblance, chap. 2.

## Nombre d'observations

- Soit  $\mathcal{P}$  un modèle pour  $X$
- $\mathbf{X}_{1:n} = (X_1, \dots, X_n)$  échantillon i.i.d. (indépendant, identiquement distribué) de même loi que  $X$ .
- modèle pour  $\mathbf{X}_{1:n}$  ?
  
- Loi jointe avec indépendance = loi produit  $P_{\theta,1:n} = P_{\theta}^{\otimes n}$
- On écrit encore (pour simplifier)  $P_{\theta}$ ,  $p_{\theta}$ , pour désigner la loi de  $\mathbf{X}_{1:n}$ .
- densité jointe : produit des densités
- donc vraisemblance d'un échantillon i.i.d. : produit des vraisemblances

$$p_{\theta}(\mathbf{x}_{1:n}) = \prod_{i=1}^n p_{\theta}(x_i)$$

1. Exemples

2. Formalisation

3. Théorie de la décision : concepts de base

Actions, fonction de coût

Risque

Exemples de coûts et risques associés.

# Actions

- Faire des stats : entreprendre une action  $a \in \mathcal{A}$  après avoir observé  $x \in \mathcal{X}$ .
- actions : produire
  - une estimation  $\hat{\theta} \in \Theta$ ,
  - un intervalle/ région  $R \subset \Theta$ ,
  - une réponse 0/1 à une question de type  $\theta \in \Theta_0$ .

$\rightarrow \mathcal{A} = \Theta/\mathcal{P}(\Theta)/\{0, 1\}$ .
- $\mathcal{A}$  : espace des actions.

## définition : procédure de décision

Une fonction (mesurable)  $\delta : \mathcal{X} \rightarrow \mathcal{A}$ .

## Fonction de coût

- hiérarchie entre les actions ? quelle procédure de décision choisir ?
- dépend des préférences de l'utilisateur, autrement de sa 'fonction de coût'.

### définition : fonction de coût

Une fonction de coût est une application

$$\begin{aligned} L : \Theta \times \mathcal{A} &\rightarrow [0, +\infty] \\ (\theta, a) &\mapsto L(\theta, a). \end{aligned}$$

$(L(\theta_0, a))$  est le "prix à payer" lorsque le vrai  $\theta$  vaut  $\theta_0$  et qu'on entreprend l'action  $a$ .

- **Idée** : classer les procédures  $\delta$  en fonction du 'comportement' de

$$L(\underbrace{\theta}_{\text{inconnu!}}, \delta(\underbrace{X}_{\text{aléatoire!}})).$$

1. Exemples

2. Formalisation

3. Théorie de la décision : concepts de base

Actions, fonction de coût

**Risque**

Exemples de coûts et risques associés.

# Risque

- **Idée** : classer les procédures  $\delta$  en fonction du ‘comportement’ de
- Simplification : considérer le “coût moyen”  $\stackrel{\text{def}}{=} \text{risque}$ .

## définition : Risque d’une procédure de décision.

Le risque d’une procédure de décision  $\delta$ , étant donné  $\theta$ , est :

$$R(\theta, \delta) = \mathbb{E}_{\theta}(L(\theta, \delta(X))) = \int_{\mathcal{X}} L(\theta, \delta(x)) \, dP_{\theta}(x).$$

1. Exemples

2. Formalisation

3. Théorie de la décision : concepts de base

Actions, fonction de coût

Risque

Exemples de coûts et risques associés.

## Estimation d'un paramètre

$g(\theta) \in \mathbb{R}$ .

- Coût quadratique :  $L(\theta, a) = (g(\theta) - a)^2$ ,

risque quadratique :  $R(\theta, \hat{g}) = \int_{\mathbb{R}} (g(\theta) - \hat{g}(x))^2 dP_{\theta}(x)$ .

- Coût  $L_1$ ,  $L(\theta, a) = |g(\theta) - a|$ ,

risque  $L_1$  :  $R(\theta, \hat{g}) = \int_{\mathbb{R}} |g(\theta) - \hat{g}(x)| dP_{\theta}(x)$ .

# Test

$a \in \{0, 1\}$ . (Question  $\theta \in \Theta_0$  ? )

$a = 0$  : "oui,  $\theta \in \Theta_0$ " ;  $a = 1$  : "non,  $\theta \notin \Theta_0$ ".

On note  $\Theta_1 = \Theta \setminus \Theta_0$ .

- coût 0 – 1 :

$$L(\theta, a) = \begin{cases} 0 & \text{si } \theta \in \Theta_a \\ 1 & \text{si } \theta \notin \Theta_a. \end{cases}$$

- risque associé :

$$R(\theta, \delta) = \begin{cases} \mathbb{P}_\theta(\delta(X) = 1) = \mathbb{E}_\theta(\delta(X)) & (\theta \in \Theta_0) \\ \mathbb{P}_\theta(\delta(X) = 0) = 1 - \mathbb{E}_\theta(\delta(X)) & (\theta \in \Theta_1) \end{cases}$$

## Région de confiance

ex :  $\Theta = \mathbb{R}$ ,  $\mathcal{A} = \{ \text{intervalles } I \subset \mathbb{R} \}$ .

- Coût 0 – 1 (encore) :

$$L(\theta, I) = \begin{cases} 0 & \text{si } \theta \in I \\ 1 & \text{si } \theta \notin I. \end{cases}$$

- risque associé :

$$R(\theta, \delta) = \mathbb{P}_\theta(\delta(X) \not\in \theta)$$

# Exemple : Prospection pétrolière

**au tableau**