Apprentissage statistique
Chapitre 1 : Modèle statistique

Lucie Le Briquer

8 janvier 2019

Table des matières

1 Exemples de modèles et problèmes statistiques

Exemple. (sondage)

On interroge n individus sur leur intention de vote entre M et P, on note y_i la réponse du $i-$ème individus, $y_i \in \{M, P\}$.

$$n_M = \sum_{i=1}^{n} \mathbb{1}\{y_i = M\} \qquad n_P = \sum_{i=1}^{n} \mathbb{1}\{y_i = P\}$$

Questions.

1. Peut-on faire une prédiction sur le candidat qui va gagner ?
2. Peut-on prédire les scores des candidats ?
3. Peut-on mesurer l'incertitude de mes réponses ?

Exemple. (reconstruction d'un signal)

On considère que l'on a une fonction $f\colon [0,1] \longrightarrow \mathbb{R}$, on a pas accès directement à f mais uniquement à sa valeur en certains points $t_1, \ldots, t_n \in [0,1]$, $t_1 < \ldots < t_n$ avec un temps d'échantillonnage T_e ($t_i = iT_e$). Au lieu d'observer $\{f(t_i)\}_{i\in\{1,\ldots,n\}}$ on observe :

$$y_i = f(t_i) + e_i$$

où e_i modélise le bruit e_i, l'erreur de mesure.

Objectif. On cherche à reconstruire f, i.e. à définir $\hat{f}\colon [0,1] \longrightarrow \mathbb{R}$ qui approxime f en un certain sens (mauvaise idée : joindre de manière linéaire les y_i).

La difficulté provient des trois caractéristiques :

- temps d'échantillonnage T_e
- $(e_i)_{i\in\{1,\ldots,n\}}$ "caractéristique" du bruit
- complexité du modèle de f

Dans le cas où f est simplement constante ou linéaire on a juste à retrouver soit sa valeur soit les coefficients.

Exemple. (contrôle de qualité avec données censurées)

On a une usine produisant des tanks, on aimerait estimer la fiabilité du tank. Pour cela, on suppose qu'on a produit n tanks. On observe pour chacun d'entre eux le premier instant de disfonctionnement $t_1, \ldots, t_n$.

Questions.

1. Quelle est la durée moyenne sans panne d'un tank ?
2. On se fixe T, on veut estimer la "probabilité" qu'un tank tombe en panne entre $[0, T]$?

En général, il est coûteux en temps d'attendre qu'un tank tombe en panne. Un autre modèle est alors le suivant : on fixe T_c et on observe les tanks tombés en panne sur $[0, T_c]$.

2 Définition formelle

2.1 Intuition

Le point de départ du statisticien est un triplet :

1. Les données, observations $(y_1, \ldots, y_n)$ qui sont à valeurs dans $(\mathbb{Y}, \mathcal{Y})$ un espace mesurable.
2. Modèles probabilistes ou statistiques. On va supposer que les (y_i) sont des réalisations de variables aléaoires $Y_1, \ldots, Y_n$ sur $(\Omega, \mathcal{F})$. Notons $z = (y_1, \ldots, y_n)$ et $Z = (Y_1, \ldots, Y_n)$. $z = Z(\omega)$ pour $\omega \in \Omega$.

   On se donne ensuite une famille de probabilités sur Ω notée $\mathcal{P}$. À partir de $\mathcal{P}$, on peut définir :
   $$\mathcal{P}^Z = \{\mathbb{P}_Z \mid \mathbb{P} \in \mathcal{P}\}$$
   Alors $\mathcal{P}^Z$ définit une famille de loi sur $\mathbb{Z} = \mathbb{Y}^n$ et $\mathcal{Z} = \mathcal{Y}^{\otimes n}$. Maintenant l'idée est de trouver une loi $\hat{\mathbb{P}}_Z \in \mathcal{P}^Z$ qui explique au mieux mes données. De manière informelle, le statisticien suppose que l'on a une vraie $\mathbb{P}_Z^*$ qui est la distribution de Z et on veut trouver $\hat{P}$ "proche" de $\mathbb{P}_Z^*$.

   Exemple. (sondage)

   On suppose que mes réponses $\in \{0, 1\}$. Un modèle dans ce cas :
   $$\mathcal{P} = \{\mathrm{Ber}(p) \mid p \in \{0, 1\}\}$$
   On suppose que mes données proviennent de $\mathrm{Ber}(p^*)$ et on veut estimer p^*.
3. Question que l'on veut résoudre.
   - reconstruction du signal
   - prédiction sur le résultat de l'élection
   - incertitude du modèle

2.2 Définition mathématique

Définition 1 (expérience statistique)

Une expérience statistique est la donnée de :

1. Un espace mesurable $(\Omega, \mathcal{F})$
2. Un espace d'observation ou d'échantillon $(\mathbb{Z}, \mathcal{Z})$, espace mesurable
3. Une variable aléatoire $Z \colon \Omega \longrightarrow \mathbb{Z}$

Définition 2 (modèle statistique)

Un modèle statistique associé à une expérience est la donnée d'une famille de lois $\mathcal{P}$ sur $(\Omega, \mathcal{F})$. On note le modèle associé :
$$\big((\Omega, \mathcal{F}), (\mathbb{Z}, \mathcal{Z}), Z, \mathcal{P}\big)$$

Définition 3 (modèle paramétrique)

On dit que le modèle est paramétrique si il existe $\Theta \subset \mathbb{R}^d$ tel que :

$$\mathcal{P} = \{\mathbb{P}_\theta \ : \ \theta \in \Theta\}$$

i.e. il existe une fonction φ surjective de Θ dans $\mathcal{P}$.
Dans l'autre cas on dit que θ est non-paramétrique.

Définition 4 (identifiable)

On dit que le modèle est identifiable s'il est paramétrique et $\varphi \colon \Theta \longrightarrow \mathbb{R}^d$ est injective.

Remarques.

1. Il sera toujours sous-entendu $\mathcal{P} = \{\mathbb{P}_\theta \ : \ \theta \in \Theta\}$ mais $\Theta \not\subset \mathbb{R}^d$ pour $d < \infty$.
2. On supposera la plupart du temps que $\Omega = \mathbb{Z}$ et $\mathcal{F} = \mathcal{Z}$ car on est juste intéressés par la famille de lois images $\mathcal{P}^Z = \{\mathbb{P}_Z \ : \ \mathbb{P} \in \mathcal{P}\}$. En général, on se prend juste $(\mathbb{Z}, \mathcal{Z})$ et $\mathcal{P}$ et on pose : $\Omega = \mathbb{Z}$, $\mathcal{F} = \mathcal{Z}$ et $Z = \mathrm{id}$.

Définition 5 (modèle statistique canonique)

Un modèle statistique canonique est juste la donnée de :

- $(\mathbb{Z}, \mathcal{Z})$ un espace mesurable
- $\mathcal{P}$ une famille de lois sur $\mathbb{Z}$

Exemple. (sondage)
On suppose que on interroge n individus parmi N, $n \ll N$. $(y_1, \ldots, y_n) \in \{0,1\}^n$

1. $$\mathbb{Z} = \{0,1\}^n \qquad Z = (Y_1, \ldots, Y_n) \qquad \mathcal{Z} = 2^{\{0,1\}^n}$$
$$\mathcal{P} = \left\{ \mathrm{Ber}(p)^{\otimes n} \ : \ p \in \frac{1}{N}\{0, \ldots, N\} \right\}$$

2. $N_1 = \sum_{i=1}^n \mathbb{1}\{Y_i = 1\}$, $N_0 = \sum_{i=1}^n \mathbb{1}\{Y_i = 0\}$.
$$\tilde{\mathbb{Z}} = \{0, \ldots, n\} \qquad \mathcal{Z} = 2^{\tilde{\mathcal{Z}}} \qquad \mathcal{P} = \left\{ \mathcal{B}(p,n) \ : \ p \in \frac{1}{N}\{0, \ldots, N\} \right\}$$

2.3 Modèles statistiques induits

Définition 6 (statistique, modèle induite)

- Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ un modèle canonique et $(\mathbb{T}, \mathcal{T})$ un espace complet et séparable $((\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)))$. On appelle statistique T toute application mesurable de $(\mathbb{Z}, \mathcal{Z})$ dans $(\mathbb{T}, \mathcal{T})$.
- On appelle modèle induit par T est le modèle :
$$(\mathbb{T}, \mathcal{T}, \mathcal{P}^T) \quad \text{où} \quad \mathcal{P}^T = \{\mathbb{P}_T \ : \ \mathbb{P} \in \mathcal{P}\}$$

Définition 7 (statistiques indépendantes)

Soient T_1, T_2 deux statistiques, elles sont indépendantes ssi :

$$\forall \mathbb{P} \in \mathcal{P},\ (T_1, T_2) \text{ indépendant suivant } \mathbb{P}$$

Définition 8 (statistiques identiquement distribuées)

Soient T_1, T_2 deux statistiques, elles sont identiquement distribuées ssi :

$$\forall \mathbb{P} \in \mathcal{P},\ \mathbb{P}_{T_1} = \mathbb{P}_{T_2}$$

Exemple. (sondage) $z \in \{0,1\}^n$, les observations (ou statistiques) canoniques sont :

$$z = (y_1, \ldots, y_n) \mapsto y_i$$

pour $i = 1, \ldots, n$. Sous le modèle donné précédemment, elles sont i.i.d.

2.4 Modèles produits et $n-$échantillon

Définition 9 (modèle produit)

Soient $(\mathbb{Z}_1, \mathcal{Z}_1, \mathcal{P}_1)$ et $(\mathbb{Z}_2, \mathcal{Z}_2, \mathcal{P}_2)$ deux modèles statistiques. On appelle modèle produit de ces deux modèles :

$$(\mathbb{Z}_1 \times \mathbb{Z}_2, \mathcal{Z}_1 \otimes \mathcal{Z}_2, \mathcal{P}_1 \otimes \mathcal{P}_2)$$

où $\mathcal{P}_1 \otimes \mathcal{P}_2 = \{\mathbb{P}_1 \otimes \mathbb{P}_2 \ :\ \mathbb{P}_1 \in \mathcal{P}_1, \mathbb{P}_2 \in \mathcal{P}_2\}$.

Remarques.

- Si on a n modèles, on notera le modèle associé :
$$\left(\prod_{i=1}^{n} \mathbb{Z}_i, \bigotimes_{i=1}^{n} \mathcal{Z}_i, \bigotimes_{i=1}^{n} \mathcal{P}_i\right)$$
- Les applications $\prod_{i=1}^{n} \mathbb{Z}_i \longrightarrow \mathbb{Z}_i$ données par $(z_1, \ldots, z_n) \mapsto z_i$ sont appelées observations/statistiques/applications canoniques.

Définition 10 ($n-$échantillon)

Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ un modèle statistique. On appelle modèle à $n-$échantillon le modèle statistique :

$$(\mathbb{Z}^n, \mathcal{Z}^{\otimes n}, \mathcal{P}^{\otimes n})$$

Lemme 1

Dans le modèle à $n-$échantillon, les observations sont i.i.d. De plus, le modèle induit par chacune d'elle est $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$.

Remarque. Pour simplifier les énoncés et la rédaction on remplace le plus souvent :

"Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ un modèle statistique, $(\mathbb{Z}^n, \mathcal{Z}^{\otimes n}, \mathcal{P}^{\otimes n})$ un $n-$échantillon et $Z_1, \ldots, Z_n$ les observations canoniques."

Par :

- *"Soit $(Z_1, \ldots, Z_n)$ i.i.d. de loi $\mathbb{P}_\theta$ pour $\theta \in \Theta$."*
- ou *"Soit $(Z_1, \ldots, Z_n)$ un $n-$échantillon de $\mathbb{P}_\theta$ pour $\theta \in \Theta$."*

Exemple. (censure)
Le premier modèle présenté "avec les mains" :

$$\mathbb{Z} = \mathbb{R}_+ \qquad \mathcal{Z} = \mathcal{B}(\mathbb{R}_+) \qquad \mathcal{P} = \{\mathrm{Exp}(\theta) \ : \ \theta \in \mathbb{R}_+^* = \Theta\}$$

Le $n-$échantillon associé à deviner. On considère $Y_1, \ldots, T_n$ les observations canoniques.

$$\tilde{Y}_i = Y_i \wedge T_c$$

On peut trouver le modèle induit par $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$.

$$\tilde{\mathbb{Z}}^n = [0, T_c]^n \qquad \tilde{\mathcal{Z}}^{\otimes n} = \mathcal{B}([0, T_c])^{\otimes n}$$

Il reste à trouver $\tilde{\mathcal{P}} = \mathcal{P}^{(\tilde{Y}_1, \ldots, \tilde{Y}_n)}$. Pour cela on remarque $(\tilde{Y}_i)$ sont i.i.d., on en déduit que $\forall \mathbb{P} \in \mathcal{P}$, $\mathbb{P}_{(\tilde{Y}_1, \ldots, \tilde{Y}_n)} = \tilde{\mathbb{P}}^{\otimes n}$ où $\tilde{\mathbb{P}} = \mathbb{P}_{\tilde{Y}_i}$.

3 Modèle statistique dominé

Définition 11

Soit $(\mathbb{X}, \mathcal{X})$ un espace mesurable et μ $\sigma-$finie.

1. Soit $f : \mathbb{S} \longrightarrow \overline{\mathbb{R}}_+$ mesurable. Alors, l'application :
$$A \mapsto \int_A f(x) d\mu(x)$$
définit une mesure sur $(\mathbb{X}, \mathcal{X})$.
2. Soit ν une mesure sur $(\mathbb{X}, \mathcal{X})$. On dit que ν admet une densité si $\exists f : \mathbb{X} \longrightarrow \mathbb{R}_+$ tel que :
$$\nu(A) = \int_A f(x) d\mu(x) \quad \forall A \in \mathcal{X}$$
f est appelée une densité de ν.
3. Si f_1 et f_2 sont deux densités de ν alors $f_1 = f_2$ $\mu-$p.p.
4. On dit que ν est absolument continue par rapport à μ si $\forall A \in \mathcal{X}$ $\mu(A) = 0$ $\nu(A) = 0$.
5. μ et ν sont dites sigulières si $\exists A \in \mathcal{X}$ tel que :
$$\mu(A) = 0 \qquad \text{et} \qquad \nu(A^C) = 0$$

Notation.

- On note $f\ d\nu/d\mu$.
- Si ν admet une densité par rapport à μ on dit que μ domine ν.

Propriété 1

Soit ν qui admet une densité par rapport à μ. Alors $\forall h\colon \mathbb{X} \longrightarrow \mathbb{R}_+$:

$$\int_{\mathbb{S}} h(x)d\nu(x) = \int_{\mathbb{X}} h(x)\frac{d\nu}{d\mu}(x)d\mu(x)$$

Théorème 1 (Radon-Nikodym)

Soit $(\mathbb{X}, \mathcal{X})$ un espace mesurable, μ $\sigma-$finie sur $(\mathbb{X}, \mathcal{X})$. Soit ν une mesure sur $(\mathbb{X}, \mathcal{X})$. Alors $\nu \ll \mu$ ssi ν admet une densité par rapport à μ : f. Et :

1. ν est $\sigma-$finie ssi $f < \infty$ $\mu-$p.p.
2. ν est finie si $\int_X f d\mu < \infty$

Remarque. On considèrera toujours des densités à valeurs dans $\mathbb{R}_+$.

Preuve. (idée) On se ramène à μ et ν finies. On considère :

$$M = \sup\left\{\int_{\mathbb{X}} f(x)d\mu(x)\ :\ f\colon \mathbb{X} \longrightarrow \mathbb{R}_+ \text{ mes. et } \int_A f(x)d\mu(x) \leqslant \nu(A)\ \forall A \in \mathcal{X}\right\}$$

On considère une suite (f_n) telle que $\int f_n d\mu \longrightarrow M$. Et on pose $f = \limsup f_n$. On a par définiton que pour tout $A \in \mathcal{X}$,

$$\int_A f(x)d\mu(x) \leqslant \nu(A)$$

Il reste à montrer que $\int_A f(x)\mu(dx) = \nu(A)$ pour tout A. Pour cela on raisonne par l'absurde. ■

Définition 12 (modèle statistique dominé)

Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ un modèle statistique. On dit qu'il est dominé par μ $\sigma-$finie sur $(\mathbb{Z}, \mathcal{Z})$ si $\forall \mathbb{P} \in \mathcal{P}$, $\mathbb{P}$ admet une densité par rapport à μ.

Remarques.

- On pourra directement considérer des modèles associés à des familles de densités par rapport à une mesure, le modèle est alors automatiquement dominé.
- Supposons que l'on a deux mesures de domination μ_1 et μ_2. On en déduit deux familles de densité $\{f_\theta^{(1)}\ :\ \theta \in \Theta\}$ et $\{f_\theta^{(2)}\ :\ \theta \in \Theta\}$. On considère $\mu_1 + \mu_2$, elle domine aussi le modèle. Par unicité des densité, on en déduit $\forall \theta \in \Theta$, $(\mu_1 + \mu_2)-$p.p. :

$$\frac{d\mathbb{P}_\theta}{d(\mu_1 + \mu_2)} = \frac{d\mathbb{P}_\theta}{d\mu_1}\frac{d\mu_1}{d(\mu_1 + \mu_2)} = \frac{d\mathbb{P}_\theta}{d\mu_2}\frac{d\mu_2}{d(\mu_1 + \mu_2)}$$

  Alors, il existe $\psi\colon \mathbb{X} \longrightarrow \overline{\mathbb{R}}_+$ tel que $\forall \theta$ $(\mu_1 + \mu_2)-$p.p. :

$$\frac{d\mathbb{P}_\theta}{d\mu_1}(x) = \frac{d\mathbb{P}_\theta}{d\mu_2}(x)\psi(x)$$

Lemme 2

Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ un modèle dominé par μ. Il existe une famille dénombrable $(\mathbb{P}_n)_{n\in\mathbb{N}} \subset \mathcal{P}$ telle que :

$$\nu = \sum_{n\geqslant 1} \frac{1}{2^n}\mathbb{P}_n \quad \text{est aussi une mesure de domination}$$

Preuve.

Comme μ est $\sigma-$finie, on peut toujours la remplacer par $\bar{\mu}$ mesure de probabilité qui domine $\mathcal{P}$ par quelque chose comme :

$$\bar{\mu} = \sum c_k \frac{\mu(\cap A_k)}{\mu(A_k)} \quad \text{avec } \sum c_k = 1$$

et A_k suite croissante d'ensembles de mesures finies. On suppose que $\mu(\mathbb{Z}) = 1$. On définit :

$$\mathcal{Q} = \left\{ \sum c_k \mathbb{P}_k \ : \ (\mathbb{P}_k) \in \mathcal{P}^{\mathbb{N}^*}, \ \sum c_k = 1 \right\}$$

$\mathcal{Q}$ est stable par combinaison convexe.

Supposons que $\mathbb{Q}^* = \sum c_k \mathbb{P}_k$ domine $\mathcal{Q}$. Soit $\nu^* = \sum \frac{1}{2^k}\mathbb{P}_k$. On suppose par l'absurde que $\exists A$ tel que $\mathbb{Q}(A) > 0$ et $\nu^*(A) = 0$. Alors $\mathbb{P}_k(A) = 0$ $\forall k$ et donc $\mathbb{Q}^*$ ne domine pas $\mathbb{Q}$.

Il suffit donc d'exhiber un $\mathbb{Q} \in \mathcal{Q}$ qui domine $\mathcal{Q}$, comme $\mathcal{P} \subset \mathcal{Q}$ on peut conclure.

1. $$\mathcal{C} = \{A \ : \ \exists \mathbb{Q} \in \mathcal{Q} \text{ t.q. } f_{\mathbb{Q}}|_A > 0 \ \mu-\text{p.p.}\}$$
   où $(f_{\mathbb{Q}})_{\mathbb{Q}\in\mathcal{Q}}$ sont définis comme $f_{\mathbb{Q}} = \frac{d\mathbb{Q}}{d\mu}$. $\mathcal{C}$ est stable par union finie. En effet, si A_1 et A_2 sont dans $\mathcal{C}$ alors $A_1 \cup A_2 \in \mathcal{C}$ en considérant $\frac{\mathbb{Q}_1+\mathbb{Q}_2}{2}$.
2. $M = \sup_{A\in\mathcal{C}} \mu(A)$. Le sup existe si $\mathcal{C}$ est non vide. C'est le cas car, si l'on suppose $\mathcal{P} \neq \emptyset$, $\mathcal{C}$ contient :
   $$A = \left(\frac{d\mathbb{P}}{d\mu}\right)^{-1}(\mathbb{R}_+^*)$$
3. On considère $(A_n)_{n\in\mathbb{N}}$ tel que $\mu(A_n) \longrightarrow M$ quand $n \to +\infty$.

Posons $\tilde{A} = \bigcup_{n\in\mathbb{N}} A_n$. Comme $\forall n \in \mathbb{N}$ $A_n \in \mathcal{C}$, $\exists \mathbb{Q}_n \in \mathcal{Q}$ tel que $f_{\mathbb{Q}_n}|_{A_n} > 0$. On définit alors :

$$\mathbb{Q}^* = \sum 2^{-n}\mathbb{Q}_n$$

On montre que $\mathbb{Q}^*$ domine $\mathbb{Q}$. $\tilde{A} \in \mathcal{C}$ (car $f_{\mathbb{Q}^*}|_{\tilde{A}} > 0$) et $\mu(\tilde{A}) = \sup_{A\in\mathcal{C}} \mu(A)$. Conséquence : $\mu|_{\tilde{A}}$ est équivalent à $\mathbb{Q}^*_{\tilde{A}}$. Soit $\mathbb{Q} \in \mathcal{Q}$ et $A \in \mathbb{Z}$ tel que $\mathbb{Q}^*(A) = 0$. Montrons que $\mathbb{Q}(A) = 0$.

1. $\mathbb{Q}(A \cap \tilde{A}) = 0$ car comme $\mu|_{\tilde{A}}$ et $\mathbb{Q}^*_{\tilde{A}}$ sont équivalentes on a $\mathbb{Q}^*(A) = 0 \Rightarrow \mathbb{Q}^*(A\cap\tilde{A}) = 0 \Rightarrow \mu(A \cap \tilde{A}) = 0$. On conclut car μ domine $\mathbb{Q}$.
2. $\mathbb{Q}(A \cap \tilde{A}^C) = 0$. Soit $B = \{f_{\mathbb{Q}} > 0\}$. Par définition il suffit de montrer que :
   $$\mathbb{Q}(A \cap \tilde{A}^C \cap B) = 0$$

On suppose par l'absurde que $\mathbb{Q}(A \cap \tilde{A}^C \cap B) > 0$. Alors comme $A \cap \tilde{A} \cap B \in \mathcal{C}$ et $\mathcal{C}$ stable par union $\tilde{A} \sqcup (\tilde{A}^C \cap B \cap C) \in \mathcal{C}$. Or si $\mathbb{Q}(\tilde{A}^C \cap A \cap B) > 0$, $\mu(\tilde{A}^C \cap A \cap B) > 0$. On aurait alors que :

$$\mu(\tilde{A}) < \mu(\tilde{A} \sqcup (\tilde{A}^C \cap B \cap A))$$

qui est absurde car $\mu(\tilde{A}) = \sup_{\mathcal{C}} \mu$.

■