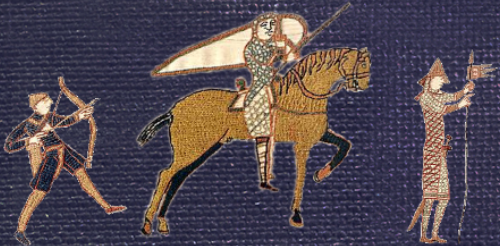
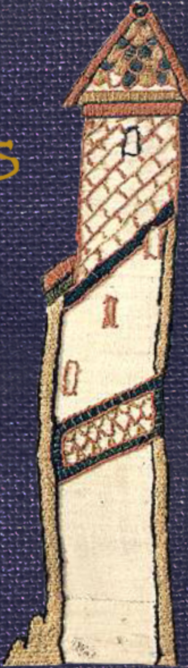
STT 1000 - STATISTIQVES

ARTHVR CHARPENTIER





Modèle paramétrique

Formalisation du problème : nous supposons disposer de $Y_1, \cdots, Y_n$ copies indépendantes d'une variable aléatoire Y dont la densité est paramétré par un paramétre réel ($\theta \in \Theta \subset \mathbb{R}$) ou vectoriel ($\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$).

On notera $\{Y_1, \cdots, Y_n\} \overset{\text{i.i.d.}}{\sim} F_\theta \in \mathcal{F}$ où $\mathcal{F}$ est la famille de lois

Exemples :

- Loi de Bernoulli $Y \sim \mathcal{B}(p)$, $\theta = p \in (0,1)$ ou $\theta = \dfrac{p}{1-p} \in \mathbb{R}_+$
- Loi de Poisson $Y \sim \mathcal{P}(\lambda)$, $\theta = \lambda \in \mathbb{R}_+$, ou $\theta = \log \lambda \in \mathbb{R}$
- Loi normale $\mathcal{N}(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$

Identifiabilité

L'application $\theta \mapsto F_\theta$ est injective,

$$\theta_1 \neq \theta_2 \Longrightarrow F_{\theta_1} \neq F_{\theta_2}$$

Modèle paramétrique

Example: Le modèle Gaussien, sur $\mathbb{R}$

$$\mathcal{F} = \left\{ f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right);\ \theta = (\mu,\sigma^2)\right\}.$$

où $\mu \in \mathbb{R}$ et $\sigma > 0$. Alors

$$f_{\theta_1} = f_{\theta_2}$$

$$\Longleftrightarrow \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right)$$

$$\Longleftrightarrow \frac{1}{\sigma_1^2}(x-\mu_1)^2 + \log\sigma_1 = \frac{1}{\sigma_2^2}(x-\mu_2)^2 + \log\sigma_2$$

$$\Longleftrightarrow x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} + \log\sigma_1 - \log\sigma_2\right) =$$

Modèle paramétrique

Example: Le modèle mélange d'exponentielles, sur $\mathbb{R}_+$

$$\mathcal{F} = \left\{ f_\theta(x) = \alpha\lambda_1 e^{-\lambda_1 x} + (1-\alpha)\lambda_2 e^{-\lambda_2 x};\ \theta = (\alpha, \lambda_1, \lambda_2) \right\}.$$

où $\alpha \in (0,1)$ et $\lambda_1, \lambda_2 > 0$.
soient $\theta_1 = (\alpha, \lambda_1, \lambda_2)$ et $\theta_2 = (1-\alpha, \lambda_2, \lambda_1)$,

$$\theta_1 \neq \theta_2 \text{ mais } f_{\theta_1} = f_{\theta_2}$$

Ce modèle n'est alors pas identifiable...

Modèle paramétrique

θ est le paramètre (en général inconnu) de la loi F_θ
Θ est l'espace des paramètres
$\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ est un échantillon aléatoire de n copies indépendantes de loi f_θ
$\boldsymbol{y} = (y_1, \cdots, y_n)$ les valeurs observées de $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$
n la taille de l'échantillon

Modèle paramétrique

Un estimateur d'un paramètre θ est une variable aléatoire (fonction de l'échantillon $\boldsymbol{Y}$) et est noté $\widehat{\theta}(\boldsymbol{Y})$. La valeur estimée de $\widehat{\theta}(\boldsymbol{Y})$ s'appelle aussi estimation et est notée $\widehat{\theta}(\boldsymbol{y})$.
(dans de nombreux ouvrages, $\widehat{\theta}$ désigne aussi bien $\widehat{\theta}(\boldsymbol{y})$ que $\widehat{\theta}(\boldsymbol{Y})$)
L'estimateur est une variable aléatoire $\widehat{\theta}(\boldsymbol{Y})$ et l'estimation est une constante $\widehat{\theta}(\boldsymbol{y})$

Example : observations suivant une loi $\mathcal{N}(\theta, 1)$

$$\widehat{\theta}_1(\boldsymbol{Y}) = \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i \text{ et } \widehat{\theta}_1(\boldsymbol{y}) = \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$$

$$\widehat{\theta}_2(\boldsymbol{Y}) = \overline{Y} = \frac{\min\{Y_i\} + \max\{Y_i\}}{2} \text{ et } \widehat{\theta}_2(\boldsymbol{y}) = \overline{y} = \frac{\min\{y_i\} + \max\{y_i\}}{2}$$

Biais

Biais d'un estimateur

On appelle biais d'un estimateur $\widehat{\theta}$ de θ la quantité

$$\text{bias}[\widehat{\theta}(\boldsymbol{Y})] = \mathbb{E}[\widehat{\theta}(\boldsymbol{Y})] - \theta$$

On dit que

- $\widehat{\theta}(\boldsymbol{Y})$ est un estimateur sans biais de θ si $\text{bias}[\widehat{\theta}(\boldsymbol{Y})] = 0$
- $\widehat{\theta}(\boldsymbol{Y})$ est un estimateur asymptotiquement sans biais de θ si

$$\lim_{n\to\infty} \text{bias}[\widehat{\theta}(\boldsymbol{Y})] = 0$$

Example $Y_1, \cdots, Y_n$ de moyenne μ,

- $\widehat{\mu}(\boldsymbol{Y}) = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ est un estimateur sans biais de μ
- $\widetilde{\mu}(\boldsymbol{Y}) = \dfrac{1}{n+3}\sum_{i=1}^{n} Y_i$ est un estimateur asymptotiquement sans biais de μ

Biais

Example $Y_1, \cdots, Y_n$ de variance σ^2,

- $\widehat{\sigma}^2(\boldsymbol{Y}) = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^n (Y_i - \overline{Y})^2$ estimate sans biais de σ^2
- $\widetilde{\sigma}^2(\boldsymbol{Y}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (Y_i - \overline{Y})^2$ est un estimateur asymptotiquement sans biais de σ^2

Biais

Example $Y_1, \cdots, Y_n$, de loi F. Soit $x \in \mathbb{R}$,

$$\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{1}(Y_i \leq y)}_{X_i}$$

où les variables X_i sont des variables de Bernoulli $\mathcal{B}(p)$ où $p = F(y)$.

$$\mathbb{E}\big[\widehat{F}(y)\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}[Y_i \leq y] = F(y)$$

donc $\widehat{F}(y)$ est un estimateur sans biais de $F(y)$, pour tout y.

Moyenne Quadratique (MSE)

Erreur Quadratique Moyenne

On appelle erreur quadratique moyenne d'un estimateur $\widehat{\theta}(\boldsymbol{Y})$ et on note $EQM(\widehat{\theta}(\boldsymbol{Y}))$ la quantité

$$EQM(\widehat{\theta}(\boldsymbol{Y})) = \mathbb{E}\left[\left(\widehat{\theta}(\boldsymbol{Y}) - \theta\right)^2\right]$$

Un peu de calcul permet d'écrire

$$EQM(\widehat{\theta}(\boldsymbol{Y})) = \text{bias}\left(\widehat{\theta}(\boldsymbol{Y})\right)^2 + \text{Var}\left(\widehat{\theta}(\boldsymbol{Y})\right)$$

Consistance

Un estimateur $\widehat{\theta}(\boldsymbol{Y})$ est consistant si

$$\lim_{n\to\infty} EQM(\widehat{\theta}(\boldsymbol{Y})) = 0$$

Moyenne Quadratique (MSE)

Pour un estimateur sans biais

$$EQM(\widehat{\theta}(\boldsymbol{Y})) = \text{Var}\big(\widehat{\theta}(\boldsymbol{Y})\big)$$

Un estimateur asymptotiquement sans biais est consistant si

$$\lim_{n\to\infty} \text{Var}\big(\widehat{\theta}(\boldsymbol{Y})\big) = 0$$

Efficacité

Soient $\widehat{\theta}_1(\boldsymbol{Y})$ et $\widehat{\theta}_2(\boldsymbol{Y})$ deux estimateurs de θ. $\widehat{\theta}_1(\boldsymbol{Y})$ est plus efficace que $\widehat{\theta}_2(\boldsymbol{Y})$ si $EQM(\widehat{\theta}_1(\boldsymbol{Y})) < EQM(\widehat{\theta}_2(\boldsymbol{Y}))$.

$$eff(\widehat{\theta}_1(\boldsymbol{Y}), \widehat{\theta}_2(\boldsymbol{Y})) = \frac{EQM(\widehat{\theta}_2(\boldsymbol{Y}))}{EQM(\widehat{\theta}_1(\boldsymbol{Y}))} \text{ rapport d'efficacité}$$

Moyenne Quadratique (MSE)

Example: $Y_1, \cdots, Y_n$ de moyenne μ,

$$\widehat{\mu}_1(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } \widehat{\mu}_2(\boldsymbol{Y}) = \frac{2}{n}\sum_{i=1}^{n/2} Y_i$$

alors $eff(\widehat{\mu}_1(\boldsymbol{Y}), \widehat{\mu}_2(\boldsymbol{Y})) = 2$

Moyenne Quadratique (MSE)

Example: $Y_1, \cdots, Y_n$ des variables $\mathcal{B}(p)$. Soit $S_n = Y_1 + \cdots + Y_n$.
$S_n \sim \mathcal{B}(n,p)$ donc $\mathbb{E}[S_n] = np$ et $\text{Var}[S_n] = np(1-p)$.

$$\widehat{p}_1 = \frac{S_n}{n} \text{ et } \widehat{p}_2 = \frac{S_n + 1}{n+2}$$

$$\mathbb{E}[\widehat{p}_1] = p \text{ et } \text{Var}(\widehat{p}_1) = \frac{p(1-p)}{n}$$

Comme c'est un estimateur sans biais de p, $EQM(\widehat{p}_1) = \dfrac{p(1-p)}{n}$

$$\mathbb{E}[\widehat{p}_2] = \frac{np+1}{n+2} \text{ et } \text{Var}(\widehat{p}_2) = \frac{\text{Var}\,(S_n)}{(n+2)^2} = \frac{np(1-p)}{(n+2)^2}$$

$$EQM(\widehat{p}_2) = \left[\frac{np+1}{n+2} - p\right]^2 + \frac{np(1-p)}{(n+2)^2} = \frac{(1-2p)^2 + np(1-p)}{(n+2)^2}$$

Moyenne Quadratique (MSE)

Aussi, le rapport d'efficacité vaut

$$eff(\widehat{p}_1, \widehat{p}_2) = \frac{EQM(\widehat{p}_2)}{EQM(\widehat{p}_1)} = \frac{n}{(n+2)^2}\left[n + \frac{(1-2p)^2}{p(1-p)}\right]$$

Si $p \sim 1/2$, ce rapport vaut $n^2/(n+2)^2 < 1$.
En fait $\widehat{p}_2$ domine $\widehat{p}_1$ si

$$p \in \left(\frac{1}{2} - \sqrt{\frac{n+1}{2n+1}}, \frac{1}{2} + \sqrt{\frac{n+1}{2n+1}}\right)$$

Exercice

Exercice: $X_1, \cdots, X_n$ de loi uniforme sur $[\theta, 2\theta]$ (avec $\theta > 0$). Montrez que l'estimateur du maximum de vraisemblance de θ est $\max\{x_i\}/2$.