

Chapitre 2: Modélisation statistique

1- Modèle statistique

Une **expérience aléatoire** est une expérience dont on ne peut pas prédire le résultat :

- lancers d'une pièce
- comptage du nb de mails reçus pendant une journée
- nombre de fois dans une main de poker

...

Une exp. aléatoire est régie par une loi de probabilité, et c'est la recherche de cette loi qui est l'objectif de la stat. inférentielle.

Pour cela il faut reproduire l'expérience aléatoire, afin de dégager des infos sur cette loi.

Le nombre d'expériences sera toujours noté n , et les résultats des expériences seront notés $x_1, x_2, \dots, x_n$. Dans ce cadre, n est la **taille de l'échantillon** et $x_1, \dots, x_n$ sont les **observations**.

L'ensemble des valeurs dans lequel se trouvent les observations s'appelle **espace des observations** : il sera noté $\mathcal{O}$, avec dans ce cours $\mathcal{O} \subset \mathbb{Z}$

Exemples

- Exp. de lancer n fois d'une pièce, en codant $\begin{cases} 1 = \text{"pile"} \\ 0 = \text{"face"} \end{cases}$, alors

$$\mathcal{O} = \{0, 1\}$$

- Exp. de comptage des mails sur n journées, alors $\mathcal{O} = \mathbb{N}$
et un exemple d'observations est $\underbrace{0, 3, 27, 2, \dots, 53}_{\text{observations sur } n \text{ jours.}}$

- On dispose d'un sac rempli de boules de couleurs rouge, bleu et vert. Répétons n fois l'exp. : on compte le nb de tirages d'une boule dans le sac, avec remise, jusqu'à obtenir une boule rouge. Des observations possibles sont $4, 1, 6, \dots, 27$ et $\mathcal{O} = \mathbb{N}^*$

Dans ce cours, les n observations $x_1, \dots, x_n \in \mathcal{O}$ sont les résultats de n expériences aléatoires identiques et indépendantes. Les résultats de ces n exp. sont alors des réalisations d'une certaine loi de probabilité $\mathcal{L}_{\theta^*}$ sur $\mathcal{O}$, pour un certain paramètre $\theta^* \in \Theta$. Autrement dit, il existe des v.a. indépendantes $X_1, \dots, X_n$ de loi $\mathcal{L}_{\theta^*}$ telles que, pour un certain $\omega \in \Omega$:

$$x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$$

Mais la seule information connue sur θ^* est son appartenance à Θ !

Donc la seule chose qu'on puisse dire est que $x_1, \dots, x_n$ sont des réalisations indépendantes de l'une des lois de la famille

$$\text{de lois } \{ \mathcal{L}_\theta \}_{\theta \in \Theta}$$

Définition (Echantillon et modèle statistique)

Un n-échantillon est une suite de v.a. $X_1, \dots, X_n$ indépendantes et identiquement distribuées, de loi $\mathcal{L}_\theta$ pour un $\theta \in \Theta$.

Le modèle statistique est l'ensemble de tous les n-échantillons.

Exemple

(1) Pour le lancer n fois d'une pièce $\begin{pmatrix} 1 = \text{"pile"} \\ 0 = \text{"face"} \end{pmatrix}$

Si $\theta = \text{probab. que la pièce tombe sur pile}$, alors chaque résultat de lancer de pièce sera une réalisation de la loi $\mathcal{B}(\theta)$.

Le modèle stat est donc décrit par l'ensemble de toutes les v.a.

$X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{B}(\theta)$, $\forall \theta \in]0, 1[$

(2) Pour l'exp. de comptage du nb de mails reçus chaque jour, et ceci pendant n jours. Nous allons montrer + bas que l'on peut considérer que la loi du nb de mails reçus chaque jour est une loi $\mathcal{P}(\theta)$, avec $\theta > 0$. Alors, le modèle stat est décrit par l'ensemble de toutes les v.a. $X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{P}(\theta)$, $\forall \theta \in \mathbb{R}_+^*$

Montrons maintenant le résultat utilisé. Notons $p \in]0, 1[$ la proba de recevoir un mail chaque minute, et k le nb de minutes dans une journée. On suppose que les mails arrivent indépendamment les uns des autres.

Alors le nb de succès reco dans une journée suit une loi $\mathcal{B}(k, p)$.

On considère $\exists \theta > 0$ t.q. $p = \frac{\theta}{k}$. Alors si $Z \sim \mathcal{B}(k, \frac{\theta}{k})$, on a

$$P(Z = \ell) = \binom{k}{\ell} \left(\frac{\theta}{k}\right)^\ell \left(1 - \frac{\theta}{k}\right)^{k-\ell} \quad \forall \ell \in \{0, \dots, k\}.$$

Calculons la limite si $k \rightarrow \infty$. On a

$$P(Z = \ell) = \frac{k(k-1) \dots (k-\ell+1)}{\ell!} \left(\frac{\theta}{k}\right)^\ell \left(1 - \frac{\theta}{k}\right)^{k-\ell}$$

Rappel

$$\ln(1+x) = x + o(x)$$

$$= \frac{1}{\ell!} \frac{k(k-1) \dots (k-\ell+1)}{k^\ell} \theta^\ell \left(1 - \frac{\theta}{k}\right)^{k-\ell}$$

$$= \frac{\theta^\ell}{\ell!} \frac{k(k-1) \dots (k-\ell+1)}{k^\ell} \exp\left((k-\ell) \ln\left(1 - \frac{\theta}{k}\right)\right)$$

$$= \frac{\theta^\ell}{\ell!} \frac{k(k-1) \dots (k-\ell+1)}{k^\ell} \exp\left((k-\ell) \left(\frac{-\theta}{k} + o\left(\frac{1}{k}\right)\right)\right)$$

$$= \frac{\theta^l}{l!} \underbrace{\frac{k(k-1)\dots(k-l+1)}{k^l}}_{\substack{\rightarrow 1 \\ k \rightarrow \infty}} \exp\left(-\underbrace{(k-l)\frac{\theta}{k}}_{\substack{\rightarrow \theta \\ k \rightarrow \infty}} + \underbrace{(k-l)o\left(\frac{1}{k}\right)}_{\substack{\rightarrow 0 \\ k \rightarrow \infty}}\right)$$

Donc $\lim_{k \rightarrow \infty} P(Z=l) = \frac{\theta^l}{l!} e^{-\theta}$ d'où Z est proche d'une loi $P(\theta)$ lorsque k est suffisamment grand. Autrement,

$$B(k, \frac{\theta}{k}) \approx P(\theta) \text{ si } k \text{ est grand}$$

les des
événements
rares

Dans notre exemple, $k = \text{nb de minutes d'un jour} = 24 \times 60$ est assez grand pour qu'on puisse considérer cette approximation.

(3) On dispose d'un sac rempli de boules rouges, bleues et vertes. On répète n fois l'exp. : on compte le nb de tirages avec remise nécessaire jusqu'à l'obtention d'une boule rouge.

Si $\theta \in]0,1[$ est la proportion de boules rouges dans le sac, la loi suivie par le nb de tirages jusqu'à obtenir une boule rouge est $\mathcal{G}(\theta)$.

Le modèle stat est donc l'ensemble de toutes les v.a. $X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{G}(\theta)$, $\forall \theta \in]0,1[$.

2. Paramètre d'intérêt et estimateur

Le **paramètre d'intérêt** est le paramètre dont on veut une approximation. Pour un modèle stat décrit par les v.a. $X_1, \dots, X_n$ indépendantes et de loi $\mathcal{L}_\theta$, $\forall \theta \in \mathcal{H}$, le paramètre d'intérêt peut être θ , ou une fonction de θ .

Dans la suite, le **paramètre d'intérêt** sera $g(\theta)$, avec $g: \mathcal{H} \rightarrow \mathbb{R}$.

Terminologie $\mathcal{H}$ s'appelle **espace des paramètres**

Le paramètre d'intérêt est la quantité que l'on veut approcher en se basant sur les observations $x_1, \dots, x_n$. En statistique, ces "approximateurs" sont appelés "estimateurs".

Définition (Estimateur) Soit $\theta \in \Theta$ et $x_1, \dots, x_n$ un n -échantillon iid de loi $\mathcal{L}_\theta$. Un estimateur de $g(\theta)$ est une v.a. $\hat{g}$ qui ne dépend que de $x_1, \dots, x_n$ et qui est à valeurs dans $g(\Theta)$.

Intérêt de cette notion Supposons que les observations $x_1, \dots, x_n \in \mathcal{O}$ sont des réalisations indépendantes de la loi $\mathcal{L}_{\theta^*}$ (θ^* inconnu).

Pour $x_1, \dots, x_n$ iid de loi $\mathcal{L}_\theta$, $\theta \in \Theta$, la vocation de l'estimateur $\hat{g} = \hat{g}(x_1, \dots, x_n)$ est d'approcher le paramètre d'intérêt $g(\theta)$, au sens où $\hat{g} \approx g(\theta)$ et ceci $\forall \theta \in \Theta$.

De ce fait, l'approximation est encore vraie pour θ^* et donc, pour les observations $x_1, \dots, x_n$ issues de la loi $\mathcal{L}_{\theta^*}$, on a donc

$$\hat{g}(x_1, \dots, x_n) \approx g(\theta^*)$$

Résumé

- On dispose de données $x_1, \dots, x_n \in \mathcal{D} \subset \mathbb{R}$, réalisations indépendantes d'une loi de probabilité $\mathcal{L}_{\theta^*}$ (sur $\mathcal{D}$)
(par ex, on lance n fois une pièce, avec le codage $1 \rightarrow \text{pile}$, $0 \rightarrow \text{face}$, on a une suite $x_1, \dots, x_n \in \{0, 1\}$ qui sont des réalisations indépendantes d'une loi $\mathcal{B}(\theta^*)$, avec $\theta^* = \text{proba que la pièce donne pile}$).
- Objectif: donner une approximation du paramètre d'intérêt $g(\theta^*)$.
- On suppose que $\theta^* \in \Theta$, l'espace des paramètres.

- On introduit le modèle statistique : c'est l'ensemble de toutes les v.a. $X_1, \dots, X_n$ qui sont indépendantes et de m.l. $\mathcal{L}_\theta$, $\theta \in \Theta$

- On construit alors un estimateur : c'est une fonction de $X_1, \dots, X_n$ v.a. indépendantes et de m.l. $\mathcal{L}_\theta$, $\theta \in \Theta$:

$$\hat{g} = \text{fonction de } X_1, \dots, X_n \text{ uniquement} = \hat{g}(X_1, \dots, X_n)$$

- Alors, si

$$\hat{g}(X_1, \dots, X_n) \approx g(\theta) \quad \underline{\underline{\forall \theta \in \Theta}}$$

On en déduit que $\hat{g}(x_1, \dots, x_n) \approx g(\theta^*)$

Pour construire des estimateurs, on utilise souvent la l. des grands nombres :

$$\mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}(X_1) \right)^2 \right] \xrightarrow{n \rightarrow \infty} 0, \text{ si } X_1, \dots, X_n \text{ v.a. indépendantes}$$

et de même l.v. D'où l'importance de $\frac{1}{n} \sum_{i=1}^n X_i$, qui mérite un nom :

Définition (Moyenne empirique) La moyenne empirique de l'échantillon $X_1, \dots, X_n$ est

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Dans la situation où la moyenne de la loi $\mathcal{L}_\theta$ est θ , $\forall \theta \in \Theta$. Si $X_1, \dots, X_n$ est un échantillon de la loi $\mathcal{L}_\theta$, alors par le LGN :

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{} \mathbb{E}(X_1) = \theta, \text{ au sens de la moyenne quadratique.}$$

$$\Downarrow \\ \mathbb{E}((\bar{X}_n - \theta)^2) \xrightarrow[n \rightarrow \infty]{} 0$$

On a donc $\bar{X}_n \approx \theta$ pour n grand, donc $\bar{X}_n$ est un estimateur "naturel" pour le paramètre d'intérêt θ .

Définition (Estimateur "plug-in", ou par insertion) Si $\hat{\theta}$ est un estimateur de θ , alors $g(\hat{\theta})$ est l'estimateur par insertion de $g(\theta)$.

Exemples

- Dans le jeu de pile ou face, avec $\text{pile} = 1$ et $\text{face} = 0$. On considère l'échantillon $X_1, \dots, X_n$ de la loi $\mathcal{B}(\theta)$, $\theta \in]0, 1[$

→ si le paramètre d'intérêt est θ .

Dans ce cas, comme $\theta = \mathbb{E}(X_1)$, et que par la LGN $\bar{X}_n \approx \mathbb{E}(X_n) = \theta$ pour n grand, $\bar{X}_n$ est un estimateur naturel pour θ .

→ si le paramètre d'intérêt est $\theta(1-\theta)$ (variance de la loi $\mathcal{B}(\theta)$)

Dans ce cas, l'estimateur par insertion est $\bar{X}_n(1-\bar{X}_n)$.

- Dans le modèle statistique décrit par l'échantillon $X_1, \dots, X_n$ i.i.d. (indépendant et identiquement distribué) de la loi $\mathcal{G}(\theta)$, avec $\theta \in]0, 1[$

On suppose que le paramètre d'intérêt est θ .

Comme $X_1 \sim \mathcal{G}(\theta)$, $\mathbb{E}(X_1) = \frac{1}{\theta}$ donc par la LGN : $\bar{X}_n \approx \frac{1}{\theta}$ si n grand.

L'estimateur par insertion de θ est alors $\frac{1}{\bar{X}_n}$