CHAPITRE 2 : MODELISATION STATISTIQUE

1. Modèle statistique

Une exp. aléatoire est une exp. dont on ne peut prédire le résultat :

- → lancer d'une pièce
- → comptage du nombre de mails reçus pendant une journée
- → nombre de rois dans une main au poker

Une exp. aléatoire est régie par une loi de proba., et sa recherche est l'objectif de la statistique inférentielle. Pour cela, on va reproduire l'exp. aléatoire.

Le nombre d'exp. aléatoires sera toujours noté n, et les résultats des exp. aléatoires seront notés $x_1, \dots, x_n$. Dans ce cadre, n est la taille de l'échantillon et $x_1, \dots, x_n$ sont les observations. L'ensemble des valeurs dans lesquelles se trouvent les observations s'appelle espace des observations. Dans ce cours, il sera noté $\mathcal{O}$, avec $\mathcal{O} \subset \mathbb{Z}$ dans ce cours dans lequel on s'intéresse à des modèles discrets.

Exemples

(1) Pour l'exp. du lancer n fois d'une pièce, on code Pile = "1" et Face = "0". Alors, les observations possibles $1, 0, 0, 1, 1, \dots$ Donc ici, $\mathcal{O} = \{0, 1\}$.

(2) On compte le nb de mails reçus chaque jour, et ceci pendant n jours. Des observations possibles sont $0, 17, 53, 4, 21, \dots$ Donc ici, $\mathcal{O} = \mathbb{N}$.

(3) On dispose d'un sac rempli de boules rouges, bleues ou vertes. 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 $4, 1, 257, 23, \dots$ Donc ici, $\mathcal{O} = \mathbb{N}^*$

Dans ce cours, les n observations $x_1, \dots, x_n \in \mathcal{O}$ sont les résultats de n exp. aléatoires identiques et indépendantes. Alors, elles sont des réalisations indépendantes d'une loi de probabilité notée $\mathcal{L}_{\theta^*}$, dépendant d'un paramètre $\theta^* \in \Theta$ (theta majuscule).

Autrement dit, il existe des v.a. indépendantes $X_1, \dots, X_n$ de loi $\mathcal{L}_{\theta^*}$ telles que, pour un $\omega \in \Omega$:

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

Mais la seule info. connue sur θ^* est son appartenance à Θ. On peut donc seulement affirmer que $x_1, \dots, x_n$ sont des réalisations indépendantes de l'une des lois de la famille de lois $\{\mathcal{L}_\theta\}_{\theta \in \Theta}$. Ceci nous amène à définir des v.a. $X_1, \dots, X_n$ indépendantes et de même loi $\mathcal{L}_\theta$, pour $\theta \in \Theta$.

Définition (Echantillon et modèle statistique)

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

Le modèle statistique est décrit par toutes ces v.a.

Exemples

(1) Pour le lancer n fois d'une même pièce. Chaque lancer de pièce est réalisation d'une loi $\mathcal{B}(\theta)$, avec $\theta \in]0,1[$. Comme les lancers sont indépendants, un n-échantillon est une suite de v.a.i.i.d. de la loi $\mathcal{B}(\theta)$, $\theta \in]0,1[$. C'est le modèle statistique de Bernoulli.

(2) On répète n fois l'exp : on tire avec remise dans un sac rempli de boules de $\neq$ couleurs jusqu'à obtenir une boule rouge. Si $\theta \in]0,1[$ est la proportion de boules rouges dans le sac, alors la loi du nombre de tirages nécessaires pour obtenir une boule rouge est $\mathcal{G}(\theta)$

Comme les tirages sont indépendants, un n-échantillon est une suite de v.a.i.i.d. de la loi $\mathcal{G}(\theta)$, pour $\theta \in]0,1[$.

2_ Paramètre d'intérêt et estimateur

Le paramètre d'intérêt est le paramètre dont on veut une approximation. Il peut être bien sûr le paramètre θ de la loi $\mathcal{L}_\theta$, mais il peut être aussi une fonction de θ.

Dans la suite, le paramètre d'intérêt sera noté $g(\theta)$, avec $g : \Theta \to \mathbb{R}$

<u>Définition</u> (Estimateur) Soient $\theta \in \Theta$ et $X_1, \dots, X_n$ un n-échantillon de la loi $\mathcal{L}_\theta$.
Un estimateur de $g(\theta)$ est une v.a. $\hat{g}$ qui ne dépend <u>que</u> de $X_1, \dots, X_n$ et qui est à valeurs dans $g(\Theta)$

Intérêt de cette notion Supposons que $x_1, \dots, x_n$ sont des réalisations de la loi $\mathcal{L}_{\theta^*}$

Pour $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{L}_\theta$, $\theta \in \Theta$, la vocation d'un estimateur $\hat{g} = \hat{g}(X_1, \dots, X_n)$ est d'approcher le paramètre d'intérêt $g(\theta)$, au sens $\overline{\infty}$ (par ex.) :

$$\mathbb{E}\left(\left(\hat{g} - g(\theta)\right)^2\right) \approx 0 \qquad \forall \theta \in \Theta$$

Comme cette propriété est vraie $\forall \theta \in \Theta$, elle est vraie aussi pour θ^* (car $\theta^* \in \Theta$)

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

Définition (Moyenne empirique) La moyenne empirique du n-échantillon $X_1, \dots, X_n$ est

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

On rappelle que (LGN) $\lim_{n \to +\infty} \overline{X}_n = \mathbb{E}(X_1)$ (au sens de la moyenne)

Par ex, si la moyenne de la loi $\mathcal{L}_\theta$ est $m(\theta)$. Alors, pour le n-échantillon $X_1, \dots, X_n$ de la loi $\mathcal{L}_\theta$, on a $\overline{X}_n \approx m(\theta)$ (si n est assez grand). De ce fait, un estimateur de $m(\theta)$ est fourni par $\overline{X}_n$.