Rappels • On dispose d'observations $x_1, \dots, x_n \in \mathcal{O} \subset \mathcal{X}$ qui sont issues de n expériences indépendantes et identiques.

• Ces n observations sont des réalisations de la loi $\mathcal{L}_{\theta^*}$, $\theta^* \in \Theta$

• θ^* inconnu $\rightarrow$ on introduit la notion de n-échantillon, qui est une suite de v.a.i.i.d $X_1, \dots, X_n$ de loi $\mathcal{L}_\theta$, pour un $\theta \in \Theta$.

2. Paramètre d'intérêt et estimateur

Le paramètre d'intérêt de l'étude stat est le paramètre dont on veut une approximation. Cela peut être θ, mais cela peut aussi être une fonction de θ. Dans toute la suite, le paramètre d'intérêt sera $g(\theta)$, avec $g : \Theta \rightarrow \mathbb{R}$.

Comment lier $g(\theta)$ aux observations $x_1, \dots, x_n$? Il faut définir une notion pour approcher $g(\theta)$, c'est un estimateur.

Définition (Estimateur) Soit $\theta \in \Theta$ et $X_1, \dots, X_n$ un échantillon i.i.d. de la 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 $x_1, \dots, x_n$ sont des réalisations indépendantes de la loi $\mathcal{L}_{\theta^*}$. Pour $X_1, \dots, X_n$ iid de loi $\mathcal{L}_\theta$, $\theta \in \Theta$, la vocation d'un estimateur $\hat{g} = \hat{g}(X_1, \dots, X_n)$ est d'être proche du paramètre d'intérêt $g(\theta)$, au sens où, par exemple

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

Cette proximité est vraie $\forall \theta \in \Theta$, donc en particulier pour θ^*. Comme les $x_1, \dots, x_n$ sont des réalisations des n var iid $X_1, \dots, X_n$ de loi $\mathcal{L}_{\theta^*}$, on en déduit

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

Pour construire des estimateurs, on utilise souvent la LGN. Elle dit que si $X_1, X_2, \ldots$ est une suite de va iid, alors

$$\lim_{n \to \infty} \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\right)^2\right) = 0.$$

Ces moyennes $\frac{1}{n}\sum_{i=1}^{n} X_i$ jouent un rôle majeur dans la construction des estimateurs. On leur donne donc un nom.

<u>Définition</u> (Moyenne empirique). La moyenne empirique de l'échantillon $X_1, \ldots, X_n$ est

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

Par exemple, si l'espérance de la loi $\mathcal{L}_\theta$ est θ, alors un estimateur du paramètre θ est ? Soit $X_1, \ldots, X_n$ un n-échantillon iid de loi $\mathcal{L}_\theta$.

Alors $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \underset{\text{LGN}}{\approx} \mathbb{E}(X_1) = \theta$, car $X_1 \sim \mathcal{L}_\theta$. Donc $\bar{X}_n$ est un estimateur naturel de θ.

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

Exemples

① Dans le modèle du jeu de pile ou face. On introduit le n-échantillon $X_1, \dots, X_n$ iid de loi $\mathcal{B}(\theta)$, $\theta \in]0,1[$. On a $\mathbb{E}(X_1) = \sum_{k=0}^{1} k\,\mathbb{P}(X_1 = k)$ $= 0 \times \mathbb{P}(X_1 = 0) + 1 \times \mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 1) = \theta$, car $X_1 \sim \mathcal{B}(\theta)$. Donc par la LGN :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \approx \mathbb{E}(X_1) = \theta \quad (\text{pour } n \text{ grand}).$$

Donc $\bar{X}_n$ est un estimateur naturel de θ. Si le paramètre d'intérêt est maintenant $\theta(1-\theta)$ (la variance de la loi $\mathcal{B}(\theta)$), alors l'estimateur (par insertion) est $\bar{X}_n(1-\bar{X}_n)$

② Dans le modèle statistique décrit par le n-échantillon $X_1, \dots, X_n$ iid de la loi $\mathcal{P}(\theta)$ avec $\theta > 0$. On a $\mathbb{E}(X_1) = \theta$ car

$$\mathbb{E}(X_1) = \sum_{k \in \mathbb{N}} k\, \mathbb{P}(X_1 = k) = \sum_{k \geq 0} k\, e^{-\theta} \frac{\theta^k}{k!} = e^{-\theta} \sum_{k \geq 0} k \frac{\theta^k}{k!}$$

$$= e^{-\theta} \sum_{k \geq 1} k \frac{\theta^k}{k!} = e^{-\theta} \sum_{k \geq 1} \frac{\theta^k}{(k-1)!}$$

$$= \theta e^{-\theta} \sum_{k \geq 1} \frac{\theta^{k-1}}{(k-1)!} \underset{\ell = k-1}{=} \theta e^{-\theta} \sum_{\ell \geq 0} \frac{\theta^\ell}{\ell!}$$

$$= \theta e^{-\theta} e^{\theta} = \theta.$$

Rappel

$$\sum_{\ell \geq 0} \frac{\theta^\ell}{\ell!} = e^\theta$$

Par la LGN, on a $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \approx \mathbb{E}(X_1) = \theta$, donc $\overline{X}_n$ est un estimateur naturel de θ.

③ Dans le modèle statistique décrit par l'échantillon $X_1, \dots, X_n$ iid de loi $\mathcal{G}(\theta)$, avec $\theta \in]0,1[$. On considère le paramètre d'intérêt θ.

On cherche un estimateur naturel pour θ. On a montré en TD que $\mathbb{E}(X_1) = \frac{1}{\theta}$ car $X_1 \sim \mathcal{G}(\theta)$. Par la LGN :

$$\bar{X}_n \approx \mathbb{E}(X_1) = \frac{1}{\theta} \quad (n \text{ grand}).$$

Donc $\frac{1}{\bar{X}_n} \approx \theta \Rightarrow \frac{1}{\bar{X}_n}$ estimateur naturel de θ.