Introduction au méthodes statistiques

Cours 2 : Inférence paramétrique

Michal W. Urdanivia*

*UGA, Faculté d'Économie, GAEL,
e-mail : michal.wong-urdanivia@univ-grenoble-alpes.fr

12 septembre 2022

1. Modèle statistique

2. Identification

3. Estimation des paramètres

4. Intervalles de confiance

Plan

1. Modèle statistique

2. Identification

3. Estimation des paramètres

4. Intervalles de confiance

1. Modèle statistique

1. Modèle statistique

- On considère le résultat observé d'une expérience statistique qui est un échantillon $X_1, X_2, \ldots, X_n$ de n v.a. i.i.d. sur un espace mesurable $(\mathcal{E}, \mathcal{F})$, avec typiquement $\mathcal{E} \subseteq \mathbb{R}$, et de distribution commune notée P.
- **Définition formelle :** un **modèle statistique** associé à cette expérience statistique est un triplet :

$$(\mathcal{E}, \mathcal{F}, (\mathrm{P}_\theta)_{\theta \in \Theta}),$$

où :

  - $(\mathcal{E}, \mathcal{F})$ est l'espace mesurable des observations ;
  - $(\mathrm{P}_\theta)_{\theta \in \Theta}$ est une famille de mesures de probabilité sur $(\mathcal{E}, \mathcal{F})$;
  - Θ est appelé l'**ensemble des paramètres**.

1. Modèle statistique

- Généralement il sera supposé que le modèle est **correctement spécifié**, i.e., défini tel que $\mathrm{P} = \mathrm{P}_\theta$, pour un $\theta \in \Theta$.
- Ce paramètre en particulier est appelé **vrai paramètre**, et il est inconnu : le but de l'expérience statistique est de l'estimer.
- On supposera aussi que $\Theta \subseteq \mathbb{R}^d$ pour $d \geq 1$. On parle alors de **modèle paramétrique**.

1. Modèle statistique

- **Exemples :**
  1. Pour n tirages de Bernoulli :

$$\Big(\{0,1\}, \mathcal{P}(\{0,1\}), (Ber(p))_{p\in(0,1)}\Big)$$

  .
  2. Si $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}xp(\lambda)$, pour un $\lambda > 0$ inconnu :

$$\Big(\mathbb{R}_+^*, \mathcal{B}(\mathbb{R}_+^*), (\mathcal{E}xp(\lambda))_{\lambda>0}\Big).$$

  3. Si $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{P}oiss(\lambda)$, pour un $\lambda > 0$ inconnu :

$$\Big(\mathbb{N}, \mathcal{P}(\mathbb{N}), (\mathcal{P}oiss(\lambda))_{\lambda>0}\Big).$$

1. Modèle statistique

4. Si $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, pour un $\mu \in \mathbb{R}$ inconnu et $\sigma^2 > 0$:

$$\left(\mathbb{R}, \mathcal{B}(\mathbb{R}), \left(\mathcal{N}(\mu, \sigma^2)\right)_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*}\right).$$

Plan

1. Modèle statistique

2. Identification

3. Estimation des paramètres

4. Intervalles de confiance

2. Identification

2. Identification

- Le paramètre θ est qualifié d'**identifié** ssi l'application $\theta \in \Theta \mapsto \mathrm{P}_\theta$ est injective, i.e.,

$$\theta \neq \theta' \Rightarrow \mathrm{P}_\theta \neq \mathrm{P}_{\theta'}$$

1. Dans tous les exemples précédents le paramètre était identifié.
2. Si $X_i = \mathbf{1}(U_i \geq 0)$, où $U_1, U_2, \ldots, U_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, pour un $\mu \in \mathbb{R}$ et $\sigma^2 > 0$, tous les deux inobservés : μ et σ^2 ne sont pas identifiés. Néanmoins μ/σ l'est.

Plan

1. Modèle statistique

2. Identification

3. Estimation des paramètres

4. Intervalles de confiance

3. Estimation des paramètres

3. Estimation des paramètres

- **Idée :** étant donné un échantillon $X_1, X_2, \ldots, X_n$ et un modèle statistique $(\mathcal{E}, \mathcal{F}, (\mathrm{P}_\theta)_{\theta \in \Theta})$, on veut estimer le paramètre θ.
- **Définitions** :
  - **Statistique** : toute fonction mesurable de l'échantillon, e.g., $\bar{X}_n$, $\max_i X_i$, $X_1 + \log(1 + |X_n|)$, variance empirique, etc, ...
  - **Estimateur** de θ : toute statistique dont l'expression ne dépend pas de θ.
  - Un estimateur $\hat{\theta}$ de θ est **faiblement consistant/convergent** ssi :

$$\hat{\theta}_n \xrightarrow{p} \theta \quad (\text{par rapport à } \mathrm{P}_\theta).$$

Remarque : lorsque la convergence est presque sûre(i.e., "$\xrightarrow{p.s.}$" à la place de "$\xrightarrow{p}$"), l'estimateur est **fortement consistant/convergent**.

3. Estimation des paramètres

- **Biais** d'un estimateur $\hat{\theta}_n$ de θ :

$$\mathrm{E}(\hat{\theta}_n) - \theta.$$

- **Risque(ou risque quadratique)** d'un estimateur $\hat{\theta}_n$:

$$\mathrm{E}\left(\left|\hat{\theta}_n - \theta\right|^2\right).$$

Remarque : si $\Theta \subseteq \mathbb{R}$,

$$\text{Risque quadratique} = \text{Biais}^2 + \text{Variance}.$$

Plan

1. Modèle statistique

2. Identification

3. Estimation des paramètres

4. Intervalles de confiance

4. Intervalles de confiance

4. Intervalles de confiance

- Soit un modèle statistique $(\mathcal{E}, \mathcal{F}, (\mathrm{P}_\theta)_{\theta\in\Theta})$ sur les observation $X_1, X_2, \ldots, X_n$, et supposons $\Theta \subseteq \mathbb{R}$.
- **Définitions** : pour $\alpha \in (0,1)$.
  - **Intervalle de confiance(C.I.) de niveau** $1-\alpha$ pour θ : tout intervalle aléatoire(i.e., dépendant de $X_1, X_2, \ldots, X_n$) $\mathcal{IC}$ dont les bornes ne dépendent pas de θ et tel que :

$$\mathrm{P}\left(\mathcal{IC} \ni \theta\right) \geq 1-\alpha, \quad \forall\theta \in \Theta.$$

  - **Intervalle de confiance(C.I.) de niveau asymptotique** $1-\alpha$ pour θ : tout intervalle aléatoire $\mathcal{IC}$ dont les bornes ne dépendent pas de θ et tel que :

$$\lim_{n\to\infty} \mathrm{P}_\theta\left(\mathcal{IC} \ni \theta\right) \geq 1-\alpha, \quad \forall\theta \in \Theta.$$

4. Intervalles de confiance

- **Exemple** : Soit $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Ber(p)$, pour un $p \in (0,1)$ inconnu.
  - LGN : la moyenne empirique $\bar{X}_n$ est un estimateur fortement consistant de p.
  - Soit t_α le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi $\mathcal{N}(0,1)$ et :

$$\mathcal{IC} = \left[\bar{X}_n - \frac{t_\alpha \sqrt{p(1-p)}}{\sqrt{n}} ; \bar{X}_n + \frac{t_\alpha \sqrt{p(1-p)}}{\sqrt{n}} \right].$$

  - TCL :

$$\lim_{n \to \infty} \mathrm{P}_p(\mathcal{IC} \ni p) = 1 - \alpha, \quad \forall p \in (0,1).$$

  - **Problème** : $\mathcal{IC}$ dépend de p !

4. Intervalles de confiance

- **Deux solutions** :
  - (i) Remplacer $p(1-p)$ par $1/4$(car $p(1-p) \leq 1/4$).
  - (ii) Remplacer p par $\bar{X}_n$ dans $\mathcal{IC}$ et utiliser le théorème de Slutsky.