CHAPITRE 15 - ESTIMATION

1 Estimation ponctuelle

1.1 Formalisme

Dans ce qui suit $(\Omega, \mathcal{A})$ est un espace probabilisable. Soit Θ une partie de $\mathbb{R}^n$, un ensemble de paramètres (en pratique $n \in \{1, 2\}$). On munit $(\Omega, \mathcal{A})$ d'une famille de probabilités P_θ avec θ décrivant Θ.

Exemples :

- $\Theta = \mathbb{R}_+^\star$ et P_θ tq $X \hookrightarrow \mathcal{E}(\theta)$.
- $\Theta = \mathbb{R} \times \mathbb{R}_+^\star$ et P_θ tq $X \hookrightarrow \mathcal{N}(\theta_1, \theta_2^2)$.

Définition : n-échantillon

On appelle n-échantillon un n-uplet $(X_1, \ldots, X_n)$ de variables i.i.d. pour tout θ.

Remarque : X_1 à X_n représentent donc le résultat de n fois la même expérience répétée de manières successives et indépendantes.

Exemples :

- On lance n fois une pièce truquée. Les $X_i \hookrightarrow \mathcal{B}(p)$ avec p la probabilité de faire pile.
- On cherche à déterminer la distribution de la durée de vies chez les personnes nées en 1932. X_1, ..., X_n peuvent être la durée de vie des n personnes prises au hasard et alors on a (vraisemblablement) $X_i \hookrightarrow N(\mu, \sigma^2)$

Remarque : Le but sera désormais d'utiliser un n-échantillon pour estimer θ ou éventuellement $g(\theta)$ une fonction de θ parfois plus facile à calculer.

Exemples :

- Pour estimer p dans le premier exemple, on peut chercher à calculer la moyenne ou la variance de la loi qui toutes deux dépendent de p.
- Idem pour la durée de vie.

Définition

Puisqu'on considère une collection de probabilités P_θ, on notera $E_\theta(X)$ l'espérance de X pour P_θ si elle existe. De même on note $V_\theta(X)$ la variance.

1.2 Estimateurs

Définition : Estimation ponctuelle et estimateur

Un estimateur de $g(\theta)$ est une variable aléatoire de la forme $T_n = \varphi(X_1, \ldots, X_n)$ où $(X_1, \ldots, X_n)$ est un n-échantillon et où $\varphi : \mathbb{R}^n \to \mathbb{R}$.
Estimer ponctuellement $g(\theta)$ par $\varphi(x_1, \ldots, x_n)$ où $(x_1, \ldots, x_n)$ est une réalisation de $(X_1, \ldots, X_n)$, c'est décider d'accorder à $g(\theta)$ la valeur $\varphi(x_1, \ldots, x_n)$.

Remarque :

> L'estimateur ne peut dépendre que de l'échantillon, à travers les valeurs des $(x_1, \ldots, x_n)$ et de sa taille. Il ne peut pas y avoir de dépendance explicite en θ.

Remarque :

> La définition ne semble pas faire référence à g ou à θ. C'est certes étonnant mais c'est à dessein : nous aurons des bons et des mauvais estimateurs. Un mauvais estimateur peut donner des valeurs complètement loufoques pour $g(\theta)$ (pourquoi pas même en dehors de $g(\Theta)$). Nous verrons comment discerner cela plus tard.

Exemples :

- Estimateurs pour p
- Estimateurs pour μ et σ.
- Voitures arrivant à un péage, estimateurs de λ
- Loi uniforme de borne sup inconnue, estimateur avec le max.

1.3 Estimateurs usuels

Définition : Moyenne empirique

On appelle moyenne empirique l'estimateur :

$$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$$

Propriétés : $E(\overline{X_n}) = \mu$ et $V(\overline{X_n}) = \frac{\sigma^2}{n}$.

Définition : Estimateur du maximum de vraisemblance

Pour $(x_1, \dots, x_n) \in \mathbb{R}^n$, soit $\hat{\theta}$ la valeur de θ qui maximise $P_\theta([X_1 = x_1] \cap [X_2 = x_2] \cap \dots \cap [X_n = x_n])$ (quantité que l'on appelle vraisemblance). Le résultat dépend de $(x_1, \dots, x_n)$ et peut s'écrire $\hat{\theta}(x_1, \dots, x_n)$.
Alors $\hat{\theta}(X_1, \dots, X_n)$ est un estimateur de θ appelé **estimateur du maximum de vraisemblance**.

Exemple : Maximum de vraisemblance pour la loi de Bernoulli $\mathcal{B}(p)$, pour la loi de Poisson $\mathcal{P}(\lambda)$.

Définition : Variance empirique

On appelle variance empirique l'estimateur :

$$\overline{S}_n = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n}.$$

1.4 Convergence d'estimateurs (HP Appli)

Définition : Estimateur convergent

Une suite d'estimateurs (T_n) de $g(\theta)$ est convergente si :

$$\forall \theta \in \Theta,\ \forall \epsilon > 0,\ P_\theta(|T_n - g(\theta)| \geq \epsilon) \xrightarrow[n \to +\infty]{} 0.$$

Exemples :

- moyenne empirique avec une loi $\mathcal{P}(\lambda)$.
- max pour la borne d'une loi uniforme $\mathcal{U}([0, a])$.
- $2X_n$ pour la borne d'une loi uniforme $\mathcal{U}([0, a])$.

2 Estimation par intervalle de confiance

Dans tout le paragraphe (U_n) et (V_n) sont des suites d'estimateurs de $g(\theta)$ telles que pour tout θ et pour tout n, $U_n \leq V_n$ presque sûrement.

2.1 Intervalle de confiance

Définition : Intervalle de confiance

Soit $\alpha \in [0, 1]$. $[U_n, V_n]$ est un intervalle de confiance de $g(\theta)$ au niveau de confiance $1 - \alpha$ si pour tout θ de Θ :

$$P_\theta([U_n \leq g(\theta) \leq V_n]) \geq 1 - \alpha.$$

Remarques :

- U_n et V_n sont des estimateurs. L'idée ici est non pas de donner une valeur de $g(\theta)$ mais de donner une borne inférieure et supérieure et de quantifier l'erreur possible.
- Explications niveau de confiance, niveau de risque

Méthode : Déterminer un intervalle de confiance

Soit T_n un estimateur de $g(\theta)$. Si $E(T_n) = g(\theta)$:

1. On écrit l'inégalité : $P\left(|T_n - g(\theta)| \geq \epsilon\right) \leq \dfrac{V(T_n)}{\epsilon^2}$.
2. On la transforme via l'événement contraire et on remarque $[|T_n - g(\theta)| < \epsilon] \subset [|T_n - g(\theta)| \leq \epsilon]$: $P\left(|T_n - g(\theta)| \leq \epsilon\right) \geq 1 - \dfrac{V(T_n)}{\epsilon^2}$.
3. On choisit ϵ pour que le membre de droite soit égale à $1 - \alpha$.
4. On transforme l'intérieur pour obtenir U_n et V_n.

Si $E(T_n) \neq g(\theta)$, on applique l'inégalité de Markov à $(T_n - g(\theta))^2$ avec $a = \epsilon^2$. Le reste fonctionne de la même manière.

Exemples :

- Intervalle de confiance pour p avec la loi $\mathcal{B}(p)$.
- Intervalle de confiance pour μ avec la loi $\mathcal{N}(\mu, \sigma^2)$ où σ^2 est connu.

2.2 Intervalle de confiance asymptotique

Définition : Intervalle de confiance asymptotique

Soit $\alpha \in [0,1]$. $([U_n, V_n])$ est un intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance $1-\alpha$ si pour tout θ de Θ, il existe une suite (α_n) à valeurs dans $[0,1]$ de limite α telle que pour tout $n \geq 1$:

$$P_\theta([U_n \leq g(\theta) \leq V_n]) \geq 1 - \alpha_n.$$

Remarque : On perd de la précision : on sait que pour n grand on a de grandes chances d'être dans l'intervalle mais on ne sait pas pour quel n.

Méthode : Déterminer un intervalle de confiance asymptotique

1. On considère $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ et $\overline{X_n}^\star = \sqrt{n}\frac{\overline{X_n}-m}{\sigma}$ où $(X_1, \ldots, X_n)$ est un n-échantillon.
2. On applique le théorème central limite pour dire que :

$$\overline{X_n}^\star \xrightarrow{\mathcal{L}} X$$

où $X \hookrightarrow \mathcal{N}(0,1)$. En particulier :

$$\lim_{n\to+\infty} P(-x \leq \overline{X_n}^\star \leq x) = 2\Phi(x) - 1.$$

3. On montre qu'il existe un unique t_α tel que $2\Phi(t_\alpha) - 1 = 1 - \alpha$ de sorte que :

$$\lim_{n\to+\infty} P(-t_\alpha \leq \overline{X_n}^\star \leq t_\alpha) = 1 - \alpha.$$

4. On inverse les arguments de la probabilité pour déterminer U_n et V_n.

Exemple : Reprendre l'exemple de l'estimation de p et comparer avec ce qui précède.
Remarque : S'il y a plus d'une quantité à déterminer (par exemple m et σ^2), l'énoncé guidera la résolution.

Proposition

Appli uniquement mais utile pour les appros.
$\left[\overline{X}_n - t_\alpha \frac{\overline{S}_n}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\overline{S}_n}{\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de μ au risque α.

3 Mathématiques approfondies

3.1 Biais des estimateurs

Définition : Biais d'un estimateur

Soit T_n un estimateur de $g(\theta)$. Si T_n admet une espérance pour tout θ, on appelle biais de T_n en $g(\theta)$ la quantité :

$$E_\theta(T_n) - g(\theta).$$

Définition : Estimateur sans biais

Soit T_n un estimateur de $g(\theta)$. On dit que T_n est sans biais si son biais est nul quelle que soit θ ou de manière équivalente si :

$$\forall \theta \in \Theta,\ E_\theta(T_n) = g(\theta).$$

Dans le cas contraire, on dit que T_n est un estimateur biaisé.

Exemples :

- Moyenne empirique
- Deux estimateurs de la variance.
- $\max(X_1, \ldots, X_n)$ pour estimer la borne sup d'une loi uniforme.

Définition : Suite d'estimateurs

Une suite d'estimateurs de $g(\theta)$ est la donnée pour tout n d'un estimateur T_n d'ordre n de $g(\theta)$. Chaque T_n est donc de la forme $\varphi_n(X_1, \ldots, X_n)$.

Définition : Suite d'estimateurs asymptotiquement sans biais

Une suite d'estimateurs (T_n) de $g(\theta)$ est dite asymptotiquement sans biais si :

$$\forall \theta \in \Theta,\ E_\theta(T_n) \xrightarrow[n\to+\infty]{} g(\theta)$$

3.2 Retour sur la convergence d'estimateurs

Définition : Estimateur convergent

Une suite d'estimateurs (T_n) de $g(\theta)$ est convergente si pour tout θ la suite (T_n) converge en probabilité vers $g(\theta)$.

Remarques :

- Les notions d'estimateurs biaisé et convergent sont distinctes. Il vaut souvent mieux un estimateur biaisé mais convergent qu'un estimateur non-biaisé et non-convergent.
- Un estimateur convergent peut tout à fait être asymptotiquement **avec** biais. Dit autrement : $X_n \xrightarrow{P} X$ n'implique pas $E(X_n) \to E(X)$.
  Exemple : $P(X_n = 0) = 1 - 1/n$ et $P(X_n = n) = 1/n$

Proposition

Si (T_n) est une suite d'estimateurs de $g(\theta)$ convergente et si $f : \mathbb{R} \to \mathbb{R}$ est continue alors $(f(T_n))$ est suite convergente d'estimateurs de $f(g(\theta))$.

Exemple : Sondage avec inversion de réponses aléatoires : au lieu de répondre directement à une question sensible, on lance un dé. Si le dé fait 1 on répond honnêtement, sinon on dit le contraire.
F_n fréquence des personnes qui répondent oui. C'est un estimateur de $q = \frac{1}{6}p + \frac{5}{6}(1-p)$.
On inverse, ça donne une fonction de F_n qui est est un estimateur de p.

Lemme

On a $E((T_n - g(\theta))^2) = V(T_n) + (E(T_n) - g(\theta))^2$.

Démonstration : *à faire* □

Proposition : Une condition suffisante de convergence

Soit (T_n) une suite d'estimateurs de $g(\theta)$. Si :

$$\forall \theta \in \Theta, \begin{cases} \lim_{n\to+\infty} E_\theta(T_n) &= g(\theta) \\ \lim_{n\to+\infty} V_\theta(T_n) &= 0 \end{cases}$$

alors (T_n) est convergente.

Démonstration : *à faire* □