Table des matières

CHAPITRE UN

PERFORMANCES

Sommaire

Dans ce chapitre nous allons étudier comment comparer des estimateurs. Un estimateur T_n de θ sera un bon estimateur s'il s'approche suffisamment du paramètre dans un sens que nous allons préciser.

1.1 Qualité d'un estimateur

1.1.1 Risque quadratique, biais et variance

Définition : *risque quadratique.*
Le risque quadratique ou erreur quadratique moyenne d'un estimateur T_n est donné par l'expression suivante :

$$EQM(T_n) = \mathbb{E}[(T_n - \theta)^2]$$

Le risque quadratique permet de comparer deux estimateurs, c'est une mesure assez brute de la qualité d'un estimateur qui peut être raffiné avec les notions de biais et de variance.

Définition : *bias d'un estimateur*
On appelle biais de T_n pour θ la valeur $b_\theta(T_n) := \mathbb{E}(T_n) - \theta$
Par conséquent on dira d'un estimateur T_n de θ qu'il est sans biais si et seulement si $\mathbb{E}(T_n) = \theta$, dans le cas contraire, on dira que T_n est un estimateur biaisé. En d'autres termes un estimateur est sans biais si son espérance est égale au paramètre recherché.

Exercice : Montrer que $EQM(T_n) = Var(T_n) + b_\theta(T_n)^2$

Le risque quadratique d'un estimateur est égal à sa variance plus le carré de son biais. Lorsqu'on évalue plusieurs estimateurs le meilleur sera sans biais et de variance minimale. Cependant, le biais et la variance ne sont pas les seuls propriété caractérisant la qualité d'un estimateur. En effet, augmenter la taille de l'échantillon est aussi à prendre en considération quand on compare des estimateurs. D'ailleurs le succès du big data revient à utiliser d'énormes échantillons, ce qui conduit à bénéficier des performances asymptotiques des estimateurs.

Exercice : On considère une variable aléatoire X telle que $\mathbb{E}[X] = \mu$ et $Var(X) = \sigma^2$.

- Montrer que les estimateurs de la moyenne et de la variance d'une loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ que vous avez trouvé à l'aide du maximum de vraisemblance sont sans biais.
- Montrer que l'estimateur de la variance empirique $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ est biaisé. Pour cela commencez par développer et simplifier cette somme.
- Déduisez en un estimateur de la variance non biaisé et retrouvez l'estimateur de la variance empirique corrigé présenté plus haut $\hat{S}_n^2$.

1.1.2 Convergence des estimateurs

Définition : *convergence d'un estimateur.*
Un estimateur T_n de θ est dit convergeant s'il converge en probabilité vers θ lorsque n tend vers l'infini. On notera $T_n \xrightarrow{\mathcal{P}} \theta$

Définition : *convergence en moyenne quadratique.*
L'estimateur T_n converge en moyenne quadratique vers θ si et seulement si son erreur quadratique moyenne tend vers 0 quand n tend vers l'infini :

$$T_n \xrightarrow{L^2} \theta \Leftrightarrow \lim_{n\to\infty} \mathbb{E}[(T_n - \theta)^2] = 0$$

Remarque : Lorsqu'un estimateur est convergent, son biais converge vers 0. On dit aussi qu'il est asymptotiquement sans biais.

Théorème *loi faible des grands nombres*
Soit $X_1, \cdots, X_n$ un ensemble de variables aléatoires *i.i.d* d'espérance finie $\mathbb{E}[X_i] = \mu$, alors l'estimateur de la moyenne empirique converge en probabilité vers la moyenne μ

$$\bar{X}_n \xrightarrow{\mathcal{P}} \mu$$

Théorème *loi forte des grands nombres*
Sous les mêmes hypothèses que le théorème précédent

$$\bar{X}_n \xrightarrow{p.s} \mu$$

Remarque :

- Les deux lois des grands nombres justifient l'interprétation fréquentiste des probabilités. C'est l'intuition selon laquelle, l'observation d'un grand nombre de données issues d'un phénomène aléatoire permet de déterminer les paramètres de la loi sous jacente. L'exemple le plus direct est le lancé d'une pièce équilibré pour obtenir pile ou face. À mesure que vous augmentez le nombre de lancés, vous obtiendrez une probabilité convergent vers 0.5. Ces théorèmes justifient théoriquement les big data pour l'estimation paramétrique.

- Vous noterez cependant que ces théorèmes concernent l'estimateur de la moyenne empirique. Cela laisse suggérer que si vous construisez votre propre estimateur à l'avenir, il faudra peut-être prouver que celui ci converge bien. Enfin, vous observerez que la seule différence entre ces deux théorèmes et la nature de la convergence sous jacente.
- Bien que la loi faible des grands nombres nous indique que la distribution de $\bar{X}_n$ se concentre autour de $\mathbb{E}[X]$, elle ne nous dit rien sur la loi que suit cet estimateur. Le théorème de la limite centrale (TCL) dit que $\bar{X}_n$ a une distribution qui est approximativement normale avec une moyenne $\mathbb{E}[X]$ et une variance $Var(X)$. Étonnamment, rien n'est supposé au sujet de la loi de X, sauf l'existence de la moyenne et de la variance.

Théorème *théorème central limite*
Soit (X_i) une suite de v.a i.i.d et $\bar{X}_n$ l'estimateur de la moyenne empirique, alors on a le résultat suivant :

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Var(X))$$

Le point clé du théorème central limite est de donner la loi de l'estimateur
Exercice :

- Montrez que les estimateurs de la moyenne empirique et de la variance empirique sont convergent en moyenne quadratique.
- Montrez l'équivalence entre les deux formulations suivantes du TCL :

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Var(X)) \Leftrightarrow \sqrt{n}(\frac{\bar{X}_n - \mathbb{E}(X)}{\sigma}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

- Donnez un diagramme d'implication comparant les 4 types de convergence suivants : convergence en loi, convergence en probabilité, convergence presque sûre, et convergence quadratique. Le but de cet exercice est de vous donner un aperçu de haut niveau des propriétés de convergence.

1.1.3 Efficacité d'un estimateur

Définition : *information de Fisher.*
On appelle quantité d'information de Fisher sur θ apporté par l'échantillon $(x_1, \cdots, x_n)$ la matrice de covariance du score :

$$I_n(\theta) = Var[s(\theta)] = Var\left[\frac{\partial \log \mathcal{L}(x_1, \cdots, x_n; \theta)}{\partial \theta}\right]$$

Proposition 1.1. *Inégalité de Cramer-Rao.*
Si T_n est un estimateur sans biais alors :

$$Var(T_n) \geq \frac{1}{I_n}$$

Ainsi une propriété de la quantité d'information de Fisher est qu'elle procure une borne inférieure à tout estimateur sans biais.

Définition : *borne de Cramer-Rao.*
La quantité $\frac{1}{I_n(\theta)}$ est appelée la borne de Cramer-Rao.
La borne de Cramer-Rao nous informe sur la variance minimale que l'on peut obtenir d'un estimateur sans biais.

Définition : *efficacité.*
On appelle efficacité d'un estimateur T_n sans biais de θ la quantité suivante :

$$Eff(T_n) = \frac{1}{I_n(\theta)Var(T_n)}$$

L'efficacité est une grandeur comprise entre 0 et 1, observez que cette quantité se rapproche de 0 lorsque la variance est très grande. Lorsque la variance est égale à la borne de Cramer-Rao, on est en présence d'un estimateur efficace. Autrement dit si $\hat{\theta}_1$ et $\hat{\theta}_2$ sont des estimateurs tels que $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$ alors $\hat{\theta}_1$ est plus efficace que $\hat{\theta}_2$. Lorsqu'un estimateur atteint l'efficacité quand n tend vers l'infini, on dit que c'est un estimateur asymptotiquement efficace. L'orsqu'un estimateur est sans biais et efficace, on dit aussi qu'il est sans biais et de variance minimale. En pratique pour comparer deux estimateurs avec un échantillon de taille n fixé on choisit celui qui à l'erreur quadratique moyenne la plus faible, i.e sans biais et efficace. Si on a accès à une infinité de données, on peut comparer leurs propriétés de convergence. Exemple, l'estimateur est-il asymptotiquement sans-biais, efficace ?

Exercice : Considérons $X_1, \cdots, X_n$ des variables aléatoires i.i.d.

1. Montrer que $I_n(\theta) = nI_1(\theta)$
2. Montrer que $I_1(\theta) = \mathbb{E}[(\nabla \log f(x;\theta))^2]$
3. Montrez la relation $\nabla^2_\theta \log(f(x;\theta)) = \frac{\nabla_\theta f(x;\theta)}{f(x;\theta)} - (\frac{\nabla_\theta f(x;\theta)}{f(x;\theta)})^2$
4. Déduisez-en cette écriture alternative $I_1(\theta) = -\mathbb{E}[\nabla^2_\theta \log \mathcal{L}(X_1;\theta)]$
5. En pratique on utilise souvent cette dernière expression pour calculer l'information de Fisher. Montrez qu'une variable aléatoire de Bernoulli de paramètre p à pour information de Fisher l'expression : $\frac{1}{p(1-p)}$.

1.1.4 Exhaustivité

On rappelle qu'un estimateur est une statistique, i.e une fonction des données, par exemple $\bar{X}_n, S^2_n$. Mais aussi $T : (X_1, \cdots, X_n) \to 1$ est une statistique.

Évidemment, il y a beaucoup de fonctions de $X_1, \cdots, X_n$ et donc beaucoup de statistiques. Lorsque nous recherchons un bon estimateur, devons-nous vraiment les considérer tous, ou existe-t-il un ensemble de statistiques beaucoup plus restreint que nous pourrions envisager ? Une autre façon de poser la question est de savoir s'il existe quelques fonctions clés de l'échantillon aléatoire qui contiennent eux-mêmes toutes les informations contenues dans l'échantillon. Par exemple, supposons que nous connaissions la moyenne de l'échantillon et la variance de l'échantillon. L'échantillon aléatoire contient-il plus d'informations sur la population que cela ? La réponse cette question dépendra de la famille de loi de probabilité que nous supposons décrire la population. Commençons par une définition heuristique d'une statistique suffisante.

Définition : *statistique exhaustive.*
Soit un modèle statistique $(\Omega, \mathcal{F}, \mathbb{P}_\Theta)$, un échantillon $X_1, \cdots, X_n$ issu de $(\Omega, \mathcal{F})$ et une statistique T. On dit que T est exhaustive si la probabilité conditionnelle de l'échantillon sachant T ne dépend pas du paramètre θ quelque soit la valeur prise par T. Autrement dit $p(X_1, \cdots, X_n | T = t)$ ne dépend pas de θ.

Le statisticien qui connaît la valeur de T peut faire un aussi bon travail d'estimation du paramètre inconnu θ que le statisticien qui connaît l'ensemble de l'échantillon aléatoire. Montrons cela avec un exercice.

Exercice : *Exemple loi Bernoulli*
Soit $X_1, \cdots, X_n$ des v.a de Bernoulli de paramètre p i.e $\mathcal{B}(p)$, montrez que $T : (X_1, \cdots, X_n) \to \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour le paramètre p.
Indices
— Commencez par exprimer $p(X_1, \cdots, X_n, T(X_1, \cdots, X_n) = t)$.
— Exprimez aussi $p(T(X_1, \cdots, X_n) = t)$.

Cependant en pratique, on ne vérifie pas directement ce critère, on utilise plutôt, le théorème de factorisation que nous présentons ci après :

Théorème 1.1. *Pour qu'une statistique T soit exhaustive pour θ, il faut et il suffit qu'il existe deux fonctions g et h mesurables telles que*

$$\forall X \in \Omega, \forall \theta \in \Theta, \ \mathcal{L}(x; \theta) = g(T(X), \theta) h(X)$$

Exercice : Trouver une statistique exhaustive pour
— une loi normale $\mathcal{N}(\mu, \sigma^2)$.
— une loi de Poisson $P(\lambda)$.
— La loi à densité suivante : $f(x|\theta) = \begin{cases} \theta x^{\theta-1} \text{ pour } x \in]0,1[\\ 0 \text{ sinon} \end{cases}$

1.2 Propriétés de l'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance n'est pas forcément unique, ni sans biais, ni efficace. Cependant, il possède d'excellentes propriétés asymptotiques.

Propriété Soit (X_i) une suite de v.a i.i.d suivant une loi P_θ, cette loi vérifiant certaines conditions de régularité, et T_n l'estimateur du maximum de vraisemblance associé à un n-échantillon. Alors on a :

— $T_n \xrightarrow{p.s} \theta$, T_n converge presque sûrement vers θ

— $\sqrt{I_n(\theta)}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, autrement dit quand n tend vers l'infini T_n est approximativement de loi $\mathcal{N}(\theta, \frac{1}{I_n(\theta)})$. On dit que T_n est asymptotiquement gaussien, sans biais et efficace. Cette propriété peut aussi s'écrire :

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{I_1(\theta)})$$

1.3 Estimation par intervalle de confiance

Dans la section précédente, on se proposait d'estimer uniquement le paramètre θ. On veut maintenant proposer un ensemble $I \subset \Theta$ aussi petit que possible, tel que θ appartienne souvent à I. Dans la suite, on suppose que X est une v.a et que $x_1, \cdots, x_n$ est un échantillon i.i.d tiré suivant la loi de X, on se donne également un estimateur T_n.

Définition : *intervalle de confiance.*
Soit $\alpha \in]0,1[$. S'il existe des v.a.r. $\theta_{min}(X_1, \cdots, X_n)$ et $\theta_{max}(X_1, \cdots, X_n)$ telles que :

$$P(\theta \in [\theta_{min}(X_1, \cdots, X_n), \theta_{max}(X_1, \cdots, X_n)] = 1 - \alpha$$

on dit alors que $[\theta_{min}(X_1, \cdots, X_n), \theta_{max}(X_1, \cdots, X_n)]$ est un intervalle de confiance pour θ, avec coefficient de sécurité $1 - \alpha$. On le note $IC_{1-\alpha}(\theta)$.

En pratique on choisit souvent $\alpha = 5\%$ ce qui emmène $1 - \alpha = 95\%$. Pour construire un intervalle de confiance du paramètre θ, on utilise les inégalités de concentration Markov, Hoeffding, Tchebychev. Ou si on connaît la loi de T_n on peut aussi directement l'utiliser. Sinon on peut utiliser une convergence en loi quand n tend vers l'infini pour construire un intervalle de confiance asymptotique.

Exemple(s) : intervalle de confiance d'une moyenne lorsque la variance est connue, lorsqu'elle est inconnue. Intervalle de confiance de la variance. Intervalle de confiance d'une proportion.