

# ***Statistiques (STA1)***

## **Cours II – Information de Fisher, estimation par maximum de vraisemblance**

---

Luca Ganassali

*Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay*

Jeudi 18 septembre 2025

**Previously in STA1...**

---

On effectue un test pour l'hypothèse nulle  $H_0$  contre l'hypothèse alternative  $H_1$  avec statistique de test  $T$ , au niveau  $\alpha$ . Après calcul, la région de rejet est choisie à  $\mathcal{R} = [0.45, 0.87]$  de sorte que  $\mathbb{P}_{H_0}(T \in \mathcal{R}) = \alpha$ .

1. La statistique de test observée vaut  $t = 0.68$ . Notre décision est de **A**: ne pas rejeter  $H_0$ ; **B**: rejeter  $H_0$ ; **C**: on ne peut pas conclure. Ici  $t = 0.68 \in \mathcal{R}$ , donc on rejette  $H_0$ . Réponse B.
2. La probabilité d'erreur de cette décision vaut **A**:  $\alpha$ ; **B**:  $t$ ; **C**:  $1 - \alpha$ ; **D**: inconnue avec les données dont on dispose. L'erreur faite ici serait de rejeter  $H_0$  a tort (erreur de première espèce), et  $\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(T \in \mathcal{R}) = \alpha$ . Réponse A.
3. Si  $t$  avait valu 0.03, la probabilité d'erreur de notre décision aurait été **A**:  $\alpha$ ; **B**:  $t'$ ; **C**:  $1 - \alpha$ ; **D**: inconnue avec les données dont on dispose. L'erreur faite ici serait de ne pas rejeter  $H_0$  a tort (erreur de seconde espèce), et  $\mathbb{P}_{H_1}(\text{ne pas rejeter } H_0) = \mathbb{P}_{H_1}(T \notin \mathcal{R})$  n'est pas calculable car on ne connaît pas  $H_1$ . Réponse D.

## **Vraisemblance et information de Fisher**

---

## Vraisemblance dans les modèles dominés

Un modèle paramétrique  $\mathcal{M} = (\mathcal{Z}, \mathbb{P}_\theta)$  est **dominé** si toutes les lois  $\mathbb{P}_\theta$  admettent une **densité**  $f_\theta$  par rapport à une mesure commune “de référence”  $\xi$  sur  $\mathcal{Z}$  (mesure de Lebesgue, mesure de comptage).

Dans un modèle paramétrique dominé, on appelle **vraisemblance** d'une réalisation  $z$  la fonction de  $\theta$ :

$$\theta \mapsto L(\theta; z) = f_\theta(z) .$$

Pour un échantillon i.i.d.,  $z = (x_1, \dots, x_n)$ , la vraisemblance s'écrit :

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i) .$$

*Remarque : bien que vraisemblance et densité aient même expression, on utilise le mot **densité** pour parler de la fonction des données à paramètre fixé (terminologie probabiliste), et le mot **vraisemblance** pour parler de la fonction du paramètre pour des données fixées (terminologie de statistiques).*

**Exemple du jour:** modèle de Bernoulli i.i.d. de paramètre  $\theta \in [0, 1]$ ,  $\mathcal{Z} = \{0, 1\}^n$ :

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = (1 - \theta)^n \prod_{i=1}^n \left( \frac{\theta}{1 - \theta} \right)^{x_i}.$$

Un modèle paramétrique  $\mathcal{M} = (\mathcal{Z}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ , dominé par  $\xi$ , et où  $\Theta$  est un ouvert de  $\mathbb{R}^d$  est **régulier** si:

- Le support des lois  $\mathbb{P}_\theta$  est indépendant de  $\theta \in \Theta$ .
- La **log-vraisemblance**  $\theta \mapsto \log L(\theta; z) =: \ell(\theta; z)$  est deux fois continûment différentiable sur  $\Theta$ , pour tout  $z \in \mathcal{Z}$ .
- Pour tout  $A$  mesurable, l'intégrale  $\int_A f(\theta; z) d\xi(z)$  est deux fois dérivable en  $\theta$  sous le signe d'intégration et on peut permuter intégration (sur  $z$ ) et dérivation (sur  $\theta$ ).

**Exemples :** échantillon Bernoulli, gaussien...

**Contre-exemple :** Loi uniforme sur  $[0, \theta]$ .

*Remarque : on n'étudiera pas ici les conditions sous lesquelles un modèle est régulier; cette propriété sera admise pour les modèles considérés (sauf indication contraire).*

Dans un modèle dominé, on appelle **fonction de score** la fonction

$$\dot{\ell}(\theta; \mathbf{z}) := \nabla_{\theta} \ell(\theta; \mathbf{z}) = \left( \frac{\partial}{\partial \theta_1} \ell(\theta; \mathbf{z}), \dots, \frac{\partial}{\partial \theta_d} \ell(\theta; \mathbf{z}) \right)^T.$$



### Proposition (Propriétés du score)

Dans un modèle paramétrique régulier,

- Le score est **additif** : pour  $Z = (X, Y)$  avec  $X, Y$  i.i.d.,

$$\dot{\ell}(\theta; x, y) = \dot{\ell}(\theta; x) + \dot{\ell}(\theta; y)$$

- Le score est un vecteur aléatoire **centré** :  $\mathbb{E}[\dot{\ell}(\theta; Z)] = \mathbf{0}$ .

En effet,

$$\begin{aligned}\mathbb{E}[\dot{\ell}(\theta; Z)] &= \mathbb{E}[\nabla_{\theta} \log L(\theta; Z)] = \mathbb{E} \left[ \frac{\nabla_{\theta} L(\theta; Z)}{L(\theta; Z)} \right] \\ &= \int \frac{\nabla_{\theta} L(\theta; z)}{L(\theta; z)} L(\theta; z) d\xi(z) = \int \nabla_{\theta} L(\theta; z) d\xi(z) \\ &= \nabla_{\theta} \underbrace{\int L(\theta; z) d\xi(z)}_{=1} = \mathbf{0}.\end{aligned}$$

Dans un modèle paramétrique régulier, on appelle **information de Fisher** au point  $\theta \in \Theta \subset \mathbb{R}^d$  la matrice de covariance du score :

$$\mathcal{I}(\theta) = \text{Var}_{\theta}(\dot{\ell}(\theta; Z)) = \mathbb{E}_{\theta}[\dot{\ell}(\theta; Z)\dot{\ell}(\theta; Z)^T].$$

C'est une matrice de taille  $d \times d$ , **symétrique définie positive**.

**Retour à l'exemple du jour:** modèle i.i.d. Bernoulli( $\theta$ ) avec  $\theta \in [0, 1]$ .

$$\ell(\theta; x_1, \dots, x_n) = n \log(1 - \theta) + (\log(\theta) - \log(1 - \theta)) \times \sum_{i=1}^n x_i$$

donc

$$\dot{\ell}(\theta; x_1, \dots, x_n) = -\frac{n}{1 - \theta} + \left(\frac{1}{\theta} + \frac{1}{1 - \theta}\right) \times \sum_{i=1}^n x_i = \frac{-n\theta + \sum_{i=1}^n x_i}{\theta(1 - \theta)}.$$

On a bien  $\mathbb{E}_{\theta}[\dot{\ell}(\theta; X_1, \dots, X_n)] = \frac{-n\theta + n\theta}{\theta(1 - \theta)} = 0$ , et

$$\mathcal{I}(\theta) = \text{Var}_{\theta}(\dot{\ell}(\theta; X_1, \dots, X_n)) = \frac{1}{\theta^2(1 - \theta)^2} \times n\theta(1 - \theta) = \frac{n}{\theta(1 - \theta)}.$$

## Proposition (Propriétés de $\mathcal{I}(\theta)$ )

Dans un modèle paramétrique régulier,

- Pour  $\ddot{\ell}(\theta; z) = \nabla^2 \ell(\theta; z)$  (hessienne, dérivée seconde), on a une **seconde expression**

$$\mathcal{I}(\theta) = \mathbb{E}_{\theta}[\dot{\ell}(\theta; Z)\dot{\ell}(\theta; Z)^T] = -\mathbb{E}_{\theta}[\ddot{\ell}(\theta; z)] .$$

- $\mathcal{I}(\theta)$  est **additive** : en notant  $\mathcal{I}_n(\theta)$  l'information de Fisher pour un  $n$ -échantillon  $Z = (X_1, \dots, X_n)$ , on a

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta) .$$

**Preuve.** 1. On a vu que

$$0 = \mathbb{E}_\theta[\dot{\ell}(\theta; Z)] = \int \nabla_\theta \ell(\theta; z) f_\theta(z) d\xi(z).$$

En dérivant encore sous l'intégrale par rapport à  $\theta$ , on obtient

$$\begin{aligned} 0 &= \int \nabla_\theta^2 \ell(\theta; z) f_\theta(z) d\xi(z) + \int \nabla_\theta \ell(\theta; z) (\nabla_\theta f_\theta(z))^T d\xi(z) \\ &= \int \ddot{\ell}(\theta; z) f_\theta(z) d\xi(z) + \int \dot{\ell}(\theta; z) \dot{\ell}(\theta; z)^T f_\theta(z) d\xi(z) \\ &= \mathbb{E}_\theta[\dot{\ell}(\theta; Z) \dot{\ell}(\theta; Z)^T] + \mathbb{E}_\theta[\ddot{\ell}(\theta; Z)]. \end{aligned}$$

2. Pour un  $n$ -échantillon indépendant  $Z = (X_1, \dots, X_n)$ , on a

$$\ell_n(\theta; Z) = \sum_{i=1}^n \ell(\theta; X_i), \quad \dot{\ell}_n(\theta; Z) = \sum_{i=1}^n \dot{\ell}(\theta; X_i).$$

Donc

$$\mathcal{I}_n(\theta) = \mathbb{E}_\theta[\dot{\ell}_n(\theta; Z) \dot{\ell}_n(\theta; Z)^T] = \sum_{i=1}^n \mathbb{E}_\theta[\dot{\ell}(\theta; X_i) \dot{\ell}(\theta; X_i)^T] = n\mathcal{I}_1(\theta),$$

où l'indépendance des  $X_i$  fait disparaître les termes croisés.

□

Intuition :  $\mathcal{I}(\theta)$  donne une idée de l'information apportée la variable aléatoire  $Z$  sur l'estimation du paramètre du modèle, i.e. la **précision** avec laquelle le paramètre peut être estimé.

### **Théorème (Borne inférieure de Cramér-Rao)**

On se place dans un modèle est paramétrique, régulier, et tel que  $\mathcal{I}(\theta)$  soit toujours inversible.

Soit  $h : \Theta \rightarrow \mathbb{R}$ . Alors, pour tout estimateur  $T$  de  $h(\theta)$ , **sans biais et de carré intégrable**, en notant  $\dot{h}(\theta) = \nabla_{\theta} h(\theta) \in$ , on a

$$\text{Var}_{\theta}(T) \geq [\dot{h}(\theta)]^T \mathcal{I}(\theta)^{-1} \dot{h}(\theta).$$

*Remarque : C'est un rapport de vitesses au carré (cf cas 1D).*

*Remarque : Attention, la borne de CR ne dit rien sur les estimateurs biaisés !*

## Borne inférieure de Cramér-Rao: preuve (cas 1D)

**Preuve dans le cas 1D** ( $\Theta = \mathbb{R}$ ). Par hypothèse,  $T$  est sans biais pour  $h(\theta)$ , donc  $\mathbb{E}_\theta[T(z) - h(\theta)] = 0$  ce qui s'écrit

$$\forall \theta \in \Theta, \int (T(z) - h(\theta))f_\theta(z)d\xi(z) = 0.$$

On dérive par rapport à  $\theta$ , en utilisant  $\frac{d}{d\theta}f_\theta(z) = \dot{\ell}(\theta; z)f_\theta(z)$ :

$$\int (T(z) - h(\theta))\dot{\ell}(\theta; z)f_\theta(z)d\xi(z) - \underbrace{\int \dot{h}(\theta)f_\theta(z)d\xi(z)}_{=\dot{h}(\theta) \times 1} = 0.$$

On met au carré et on applique Cauchy-Schwarz:

$$\begin{aligned}(\dot{h}(\theta))^2 &= \left( \int (T(z) - h(\theta))\dot{\ell}_\theta(z)f_\theta(z)d\xi(z) \right)^2 \\&\leq \left( \int (T(z) - h(\theta))^2 f_\theta(z)d\xi(z) \right) \left( \int (\dot{\ell}_\theta(z))^2 f_\theta(z)d\xi(z) \right) \\&= \text{Var}_\theta(T) \times \mathcal{I}(\theta).\end{aligned}$$

□

Un estimateur sans biais  $T(Z)$  est dit **efficace** pour estimer  $h(\theta)$  s'il atteint la borne de Cramér-Rao, i.e. si pour tout  $\theta \in \Theta$ ,

$$\text{Var}_{\theta}(T) = [\dot{h}(\theta)]^T \mathcal{I}(\theta)^{-1} \dot{h}(\theta).$$

On dit qu'il est **Uniformément de Variance Minimale** parmi les estimateurs sans **Biais** (**UVMB** ou **UMVE** en anglais).

**Retour à l'exemple du jour:** modèle i.i.d. Bernoulli( $\theta$ ) avec  $\theta \in [0, 1]$ . On avait calculé  $\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}$ . L'estimateur  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  pour  $h(\theta) = \theta$  est sans biais. Sa variance vaut  $\frac{\theta(1-\theta)}{n} = \mathcal{I}^{-1}(\theta)$  : il est efficace.

*Remarque : il peut ne pas exister d'estimateurs efficaces, et il peut y avoir des estimateurs sans biais optimaux (UVMB) non efficaces*

*Remarque : on peut s'intéresser néanmoins à la construction d'estimateurs sans biais qui atteignent asymptotiquement la borne de Cramér-Rao, quand la taille d'échantillon  $n$  tend vers  $+\infty$ .*

## **Estimation par maximum de vraisemblance**

---



On se place dans le cas d'un  $n$ – échantillon  $Z = (X_1, \dots, X_n)$ . La **méthode des moments** pour construire un estimateur de  $h(\theta)$  consiste à :

- écrire  $h(\theta)$  sous la forme  $h(\theta) = g(m_1, \dots, m_k)$  où  $m_\ell = \mathbb{E}_\theta[X_1^\ell]$ .
- remplacer les  $m_\ell$  par leurs estimateurs empiriques dans la formule:

$$\hat{h}(\theta) = g(\hat{m}_1, \dots, \hat{m}_k), \quad \text{où } \hat{m}_\ell = \frac{1}{n} \sum_{i=1}^n X_i^\ell.$$

# Méthode du maximum de vraisemblance

On appelle **estimateur du maximum de vraisemblance**, une valeur  $\theta$  maximisant la (log-)vraisemblance :

$$\hat{\theta}_{MV} \in \arg \max_{\theta \in \Theta} L(\theta; Z) = \arg \max_{\theta \in \Theta} \ell(\theta; Z).$$

**Retour à l'exemple du jour:** modèle i.i.d. Bernoulli( $\theta$ ) avec  $\theta \in [0, 1]$ . On avait calculé

$$\ell(\theta; Z) = \log(1 - \theta) \times \left( n - \sum_{i=1}^n X_i \right) + \log(\theta) \times \sum_{i=1}^n X_i,$$

de dérivée

$$\dot{\ell}(\theta; Z) = -\frac{1}{1 - \theta} \times \left( n - \sum_{i=1}^n X_i \right) + \frac{1}{\theta} \times \sum_{i=1}^n X_i,$$

et

$$\dot{\ell}(\theta; Z) = 0 \iff \theta \times \left( n - \sum_{i=1}^n X_i \right) = (1 - \theta) \times \sum_{i=1}^n X_i \iff \theta = \frac{1}{n} \sum_{i=1}^n X_i.$$

d'où  $\hat{\theta}_{MV} = \frac{1}{n} \sum_{i=1}^n X_i$ .

Pourquoi l'EMV ? Dans le cas d'un  $n$ -échantillon, on a

$$\ell_n(\theta; \mathbf{z}) = \sum_{i=1}^n \log f_{\theta}(x_i)$$

et en notant  $\theta^*$  le vrai paramètre, par la LGN, pour tout  $\theta$ , quand  $n \rightarrow \infty$ ,

$$\frac{1}{n} \ell_n(\theta; \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n \log f_{\theta}(x_i) \longrightarrow \mathbb{E}_{\theta^*} [\log f_{\theta}(X)] =: F(\theta).$$

Pour tout  $\theta$ ,

$$\begin{aligned} F(\theta^*) - F(\theta) &= \int (\log f_{\theta^*}(x) - \log f_{\theta}(x)) f_{\theta^*}(x) d\xi(x) \\ &= \int \log \frac{f_{\theta^*}(x)}{f_{\theta}(x)} f_{\theta^*}(x) d\xi(x) = \int \left( \log \frac{f_{\theta^*}(x)}{f_{\theta}(x)} \times \frac{f_{\theta^*}(x)}{f_{\theta}(x)} \right) f_{\theta}(x) d\xi(x) \\ &= \mathbb{E}_{\theta^*} \left[ \log \frac{f_{\theta^*}(X)}{f_{\theta}(X)} \times \frac{f_{\theta^*}(X)}{f_{\theta}(X)} \right] \geq \log \mathbb{E}_{\theta^*} \left[ \frac{f_{\theta^*}(X)}{f_{\theta}(X)} \right] \times \mathbb{E}_{\theta^*} \left[ \frac{f_{\theta^*}(X)}{f_{\theta}(X)} \right] = 0 \end{aligned}$$

Donc  $\theta^*$  est un maximum global de la fonction  $F$ , cela motive le choix de l'EMV.

## Inconvénients :

- Si la vraisemblance n'est pas strictement concave pour tout  $\theta$ , il peut exister des optima locaux.
- L'EMV n'est pas forcément unique.
- L'EMV peut ne pas exister.
- On peut avoir des problèmes de dérivabilité dans des modèles dominés non réguliers, par ex  $\text{Unif}(0, \theta)$ .

## Avantages:

- Pour toute bijection  $g$  de  $\Theta$  dans  $\Theta'$  (reparamétrisation), si  $\hat{\theta}$  est l'EMV de  $\theta$ , alors  $g(\hat{\theta})$  est l'EMV de  $g(\theta)$ . L'EMV est équivariant par reparamétrisation bijective.
- De bonnes propriétés asymptotiques dans les modèles ayant suffisamment de conditions de régularité.

**Proposition (consistance forte de l'EMV)** On se place dans le cas d'un  $n$ -échantillon, avec les  $X_i$  de même densité  $f_\theta$ . On suppose que

- (i) le modèle est identifiable;
- (ii)  $\Theta$  est compact et pour tout  $x \in \mathcal{X}$ ,  $\theta \rightarrow f_\theta(x)$  est continue.
- (iii) :  $h$  est dans  $L_1(\mathbb{P}_\theta)$  pour tout  $\theta$ , avec  $h : x \mapsto \sup_{s \in \Theta} |\log f_s(x)|$ .

Alors,  $\hat{\theta}_{MV}$  est fortement consistant.

*Remarque : en pratique, on fera les choses à la main, sans utiliser ce théorème.*

### Proposition (asymptotique de l'EMV)

On se place dans le cas d'un  $n$ -échantillon, avec les  $X_i$  de même densité  $f_\theta$ . Soit  $\hat{\theta}_{MV}$  l'EMV du paramètre  $\theta$ . Sous des conditions de régularité du modèle (identifiabilité, convexité ou compacité, uniformité), et si  $\mathcal{I}_1(\theta)$  est inversible, alors pour tout  $\theta \in \Theta$ , on a la convergence en loi suivante :

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \longrightarrow \mathcal{N}(\mathbf{0}, \mathcal{I}_1(\theta)^{-1}).$$

*Remarque : là encore, dans nos cas (simples), on passera plutôt par le TLC*

*Remarque :  $\hat{\theta}_{MV}$  est asymptotiquement sans biais, mais est en général biaisé pour  $n$  fini.*

*Remarque : dans le cas de la proposition ci-dessus, on parle alors d'efficacité asymptotique. Bien que la variance corresponde avec la borne de C-R, à strictement parler cette dernière ne s'applique pas à l'EMV à cause du biais, même tendant vers 0.*

Merci !

Rdv en TD pour les questions et la pratique de ces notions.

(contenu du cours disponible sur ma page web : [lganassali.github.io](https://lganassali.github.io))