

## Corrigé feuille 7 : Estimateurs et intervalles de confiance

### Exercice 1 :

Soit  $(X_1, \dots, X_n)$  un échantillon de taille  $n \geq 2$  de loi normale  $(m, \sigma^2)$  avec  $m \in \mathbb{R}$  et  $\sigma > 0$ .

1. On suppose  $m$  et  $\sigma$  inconnus. On considère les estimateurs  $T_1$  et  $T_2$  de  $\sigma^2$ . Voyons si les estimateurs sont biaisés ou pas.

$$\begin{aligned}\mathbb{E}_\sigma[T_1] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\sigma[(X_i - \bar{X}_n)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\sigma[X_i^2] - \frac{2}{n} \sum_{i=1}^n \mathbb{E}_\sigma[X_i \bar{X}_n] + \mathbb{E}_\sigma[\bar{X}_n^2] \quad \text{par linéarité de l'espérance} \\ &= \sigma^2 + m^2 - \frac{2}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_\sigma[X_i X_j] + \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_\sigma[X_i^2] + \frac{1}{n^2} \sum_{i \neq j} \mathbb{E}_\sigma[X_i X_j] \\ &= \sigma^2 + m^2 - \frac{2}{n^2} (n(\sigma^2 + m^2) + (n^2 - n)m^2) + \frac{1}{n} (\sigma^2 + m^2) + \frac{n^2 - n}{n^2} m^2 \\ &\quad \text{car } \mathbb{E}_\sigma[X_i X_j] = \sigma^2 + m^2 \text{ si } i = j, m^2 \text{ si } i \neq j \\ &= \frac{n-1}{n} \sigma^2.\end{aligned}$$

Ainsi,  $\mathbb{E}_\sigma[T_1] \neq \sigma^2$  donc  $T_1$  est un estimateur biaisé.

Comme  $T_2 = \frac{n}{n-1} T_1$ , on en déduit que  $\mathbb{E}_\sigma[T_2] = \sigma^2$ . Par conséquent,  $T_2$  est un estimateur sans biais.

Dans le cas où l'on privilégie les estimateurs sans biais, on utilisera donc  $T_2$  plutôt que  $T_1$ .

2. On suppose  $m$  connu et  $\sigma$  inconnu. On considère un nouvel estimateur  $T_3$  de  $\sigma^2$ . Montrons que  $T_3$  est sans biais.

$$\begin{aligned}\mathbb{E}_\sigma[T_3] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\sigma[(X_i - m)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\sigma[X_i^2] - \frac{2}{n} \sum_{i=1}^n m \mathbb{E}_\sigma[X_i] + \frac{1}{n} \sum_{i=1}^n m^2 \\ &= \sigma^2 + m^2 - 2m^2 + m^2 \\ &= \sigma^2.\end{aligned}$$

Aussi,  $T_3$  est un estimateur sans biais.

Le choix est maintenant à faire entre  $T_2$  et  $T_3$  qui sont tous les deux sans biais. Un critère pour les départager est de calculer les risques quadratiques et de conserver l'estimateur de risque minimal. En prenant un tout petit peu de recul, on se rend compte que dans l'estimateur  $T_2$  on considère une estimation de la moyenne alors que dans l'estimateur  $T_3$  on prend la « vraie » valeur  $m$ . Il y a donc de fortes chances que  $T_3$  soit un « meilleur » estimateur que  $T_2$ .

**Exercice 2 :**

Un estimateur  $T_n$  ( $n$  est la taille de l'échantillon utilisé) d'un paramètre  $\theta \in \mathbb{R}$  a la loi :

$$\mathbb{P}[T_n = \theta] = 1 - \frac{1}{n} ; \quad \mathbb{P}[T_n = n] = \frac{1}{n}.$$

1. Montrons que cet estimateur n'est pas convergent (vers  $\theta$ ) en moyenne quadratique. On remarque que  $\mathbb{P}[T_n - \theta = 0] = 1 - \frac{1}{n}$  et  $\mathbb{P}[T_n - \theta = n - \theta] = \frac{1}{n}$ . Alors,  $\mathbb{E}[(T_n - \theta)^2] = \frac{(n-\theta)^2}{n}$ .  $(\mathbb{E}[(T_n - \theta)^2])_n$  ne converge donc pas vers 0 quand  $n$  tend vers l'infini. Par conséquent,  $T_n$  ne converge pas vers  $\theta$  en moyenne quadratique.

2. Montrons qu'il est convergent en probabilité. Pour tout  $\varepsilon > 0$ , pour  $n$  assez grand, on a

$$\mathbb{P}[|T_n - \theta| > \varepsilon] = \mathbb{P}[T_n - \theta = n - \theta] = \frac{1}{n} \xrightarrow{n \rightarrow +\infty} 0.$$

Aussi,  $T_n$  converge vers  $\theta$  en probabilité.

**Exercice 3 :**

Soit  $(X_1, \dots, X_n)$  un échantillon de taille  $n$  de loi de Bernoulli de paramètre  $p$ . On considère l'estimateur  $T_n = \bar{X}_n(1 - \bar{X}_n)$  pour le paramètre  $\theta = p(1 - p)$ .

1. D'après la loi forte des grands nombres,  $(\bar{X}_n)$  converge presque sûrement donc en probabilité vers  $p$ . Par théorèmes d'opérations sur les limites,  $(T_n)$  converge donc en probabilité vers  $p(1 - p)$ .

2. On montre que  $\mathbb{E}_p[T_n] = \frac{n-1}{n}p(1 - p)$ . Donc  $T_n$  est un estimateur biaisé de  $p(1 - p)$ .

3. On en déduit que  $\frac{n}{n-1}T_n$  est un estimateur non biaisé de  $\theta$ .

**Exercice 4 :**

On pose  $T_1 = \frac{1}{4}$ ,  $T_2 = \frac{3}{4}$  et  $T_3 = \frac{3}{4}\mathbf{1}_{\{S=0\}} + \frac{1}{4}\mathbf{1}_{\{S=1\}} + \frac{1}{4}\mathbf{1}_{\{S=2\}}$ .

$T_1$  et  $T_2$  sont des fonctions constantes donc mesurables. D'autre part, comme  $S$  est une variable aléatoire,  $\mathbf{1}_{\{S=0\}}$ ,  $\mathbf{1}_{\{S=1\}}$  et  $\mathbf{1}_{\{S=2\}}$  sont mesurables et, par conséquent,  $T_3$  est mesurable. Ainsi,  $T_1$ ,  $T_2$  et  $T_3$  sont des estimateurs.

Calculons les risques minimax associés à ces estimateurs.

On a  $R_{T_1}(p) = \mathbb{E}[(T_1 - p)^2] = (\frac{1}{4} - p)^2$ . Donc

$$\sup_{p \in \{1/4, 3/4\}} R_{T_1}(p) = \frac{1}{4}.$$

De même,  $R_{T_2}(p) = (\frac{3}{4} - p)^2$ . Donc

$$\sup_{p \in \{1/4, 3/4\}} R_{T_2}(p) = \frac{3}{4}.$$

Enfin,

$$\begin{aligned} R_{T_3}(p) &= \mathbb{E}[(T_3 - p)^2] \\ &= \frac{9}{16}\mathbb{P}[S = 0] + \frac{1}{16}\mathbb{P}[S = 1] + \frac{1}{16}\mathbb{P}[S = 2] + p^2 \\ &\quad - 2p \left[ \frac{3}{4}\mathbb{P}[S = 0] + \frac{1}{4}\mathbb{P}[S = 1] + \frac{1}{4}\mathbb{P}[S = 2] \right]. \end{aligned}$$

De plus,  $S$  suit la loi binômiale de paramètres 2 et  $p$ . Après calculs, on trouve

$$R_{T_3}(p) = \frac{9}{16} - \frac{5p}{2} + \frac{7p^2}{2} - p^3.$$

En particulier,  $R_{T_3}(\frac{1}{4}) = \frac{9}{64}$  et  $R_{T_3}(\frac{3}{4}) = \frac{15}{64}$ . Donc

$$\sup_{p \in \{1/4, 3/4\}} R_{T_3}(p) = \frac{15}{64} < \frac{1}{4}.$$

Finalement,  $T_3$  est l'estimateur de risque minimax parmi  $T_1$ ,  $T_2$  et  $T_3$ .

**Exercice 5 : Estimateur du maximum de vraisemblance**

On considère une variable aléatoire discrète  $X$  dont la loi de probabilité est définie par

$$\mathbb{P}(X = k) = \frac{\theta^{k-1}}{(1 + \theta)^k} \text{ pour } k \in \mathbb{N}^* = \{1, 2, 3, \dots\}.$$

Ici  $\theta$  est un paramètre inconnu et strictement positif.

1. On détermine la fonction génératrice de  $X$ .

$$G_X(s) = \sum_{k=1}^{+\infty} \mathbb{P}[X = k]s^k = \theta^{-1} \sum_{k=1}^{+\infty} \left( \frac{\theta s}{1 + \theta} \right)^k = \frac{s}{\theta + 1 - \theta s}.$$

On en déduit l'espérance et la variance de  $X$  :

$$\mathbb{E}[X] = G'_X(1) = \theta + 1 \text{ et } \text{Var}(X) = G''_X(1) + G'_X(1) - G'_X(1)^2 = \theta(1 + \theta).$$

2. On s'intéresse à l'estimateur du maximum de vraisemblance du paramètre  $\theta$ .

a. On détermine d'abord la fonction de vraisemblance.

$$\begin{aligned} h(\theta, x) &= \mathbb{P}[(X_1, \dots, X_n) = (x_1, \dots, x_n)] \\ &= \prod_{i=1}^n \mathbb{P}[X_i = x_i] \text{ par indépendance} \\ &= \prod_{i=1}^n \frac{\theta^{x_i-1}}{(1 + \theta)^{x_i}} \\ &= \frac{\theta^{n(\bar{x}-1)}}{(1 + \theta)^{n\bar{x}}} \text{ où } \bar{x} = \sum_{i=1}^n \frac{x_i}{n}. \end{aligned}$$

Ensuite, l'estimateur du maximum de vraisemblance est défini de la façon suivante :

$$\hat{\theta}_n = \underset{\theta > 0}{\text{Argmax}} h(\theta, X).$$

En calculant la dérivée partielle de  $h$  par rapport à  $\theta$ , on remarque alors qu'elle est positive pour  $\theta < \bar{x} - 1$ , nulle en  $\bar{x} - 1$  et négative pour  $\theta > \bar{x} - 1$ . Aussi, l'estimateur du maximum de vraisemblance  $\hat{\theta}_n$  du paramètre  $\theta$  vaut  $\bar{X} - 1$ .

b.  $\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[\bar{X}] - 1 = \mathbb{E}[X] - 1 = \theta$ , donc  $\hat{\theta}_n$  est un estimateur sans biais de  $\theta$ .

c. D'après la loi des grands nombres,  $\bar{X}_n$  converge presque sûrement et dans  $L^2$  vers  $\theta + 1$ . Par conséquent,  $\hat{\theta}_n$  converge p.s. et dans  $L^2$  vers  $\theta$ .

**Exercice 6 :**

Soit  $m$  la durée de vie moyenne d'un composant. Soit  $X_i$  la durée de vie d'un composant. On suppose que  $X_i$  suit la loi  $\mathcal{N}(m, 70^2)$  et que les variables  $X_i$  sont indépendantes. On estime  $m$  par  $\bar{x}=450$  pour  $n = 250$ .

$\sqrt{n}\frac{\bar{X}-m}{70}$  suit la loi normale centrée réduite. On cherche un intervalle de confiance de niveau 0.99. Notant  $Y$  une variable aléatoire de loi  $\mathcal{N}(0, 1)$ , d'après les tables, on a :

$\mathbb{P}(|Y| > 2.576) = 0.01$ . Alors

$$\mathbb{P}\left(\sqrt{250}\frac{\bar{X} - m}{70} \in [-2.58, 2.58]\right) \approx 0.99.$$

On peut donc prendre l'intervalle de confiance suivant : [438.57,461.42].

**Exercice 7 :**

Soit  $X_i$  de loi de Bernoulli de paramètre  $p$  ( $X_i$  vaut 1 si la pièce est défectueuse, 0 sinon). D'après le théorème central limite, pour  $n$  assez grand, on peut approcher la loi de  $\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{p(1-p)}}$  par une loi normale centrée réduite. Ici,  $n = 200$ . De plus, d'après les tables si  $Y$  suit la loi  $\mathcal{N}(0, 1)$ , on a  $\mathbb{P}(|Y| > 2.576) = 0.01$  et  $\mathcal{P}(|Y| > 1.96) = 0.05$ . Pour trouver un intervalle de confiance de niveau 0.99, il reste donc à résoudre l'inéquation :

$$\left|\sqrt{200}\frac{1/10 - p}{\sqrt{p(1-p)}}\right| \leq 2.576.$$

En passant au carré et en résolvant l'équation du second degré en  $p$ , on trouve l'intervalle de confiance [0.0576,0.1681]. On raisonne de manière analogue dans l'autre cas.