Chapitre 23 : Estimation

1. Notion d'estimateur

1.1 Principe général

Soit X une variable aléatoire. On suppose que cette loi dépend d'un **paramètre θ** (loi de Bernoulli de paramètre p, loi de Poisson de paramètre λ…), qui est **inconnu**.
On suppose que θ appartient à un ensemble Θ.

Soit $(X_n)_{n \geq 1}$ une suite de **variables indépendantes** et qui suivent toutes la même loi que X.
(On répète par exemple l'expérience de manière indépendante)
Si $n \geq 1$, $(X_1, \ldots, X_n)$ est appelé un **n-échantillon**

On va essayer d'**estimer la valeur de θ**, à l'aide des valeurs prises par l'échantillon $(X_1, X_2, \ldots, X_n)$.

Définition :

Avec les hypothèses précédentes, on appelle **estimateur de θ**, toute variable aléatoire
$T_n = f(X_1, \ldots, X_n)$, où f est une application de $\mathbb{R}^n$ dans $\mathbb{R}$, telle que T_n est une variable aléatoire.

Exemples :
$T_n = \dfrac{X_1 + \ldots + X_n}{n}$, $T_n = \dfrac{X_{n-1} + X_n}{2}$, $T_n = \text{Max}(X_1, \ldots, X_n)$, $T_n = (X_1 X_2 \ldots X_n)^{1/n}$ (si $X_i \geq 0$) sont des estimateurs de θ.

Remarque :
Pour obtenir un bon estimateur de θ, les valeurs prises par T_n doivent être proches de θ.
Pour cela on veillera à ce que :
_ l'espérance de T_n soit égale à θ, ou au moins tende vers θ quand n tend vers + ∞
_ la variance de T_n soit la plus petite possible, et tende vers 0 quand n tend vers + ∞.

1.2 Exemple d'étude d'un estimateur : **la moyenne empirique**

Soit $\lambda > 0$. Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires indépendantes qui suivent toutes la loi de Poisson de paramètre λ. Pour $n \geq 1$, on pose $T_n = \dfrac{X_1 + \ldots + X_n}{n}$.

1) Montrer que T_n est un estimateur de λ.
2) Déterminer $E(T_n)$ et $V(T_n)$. Cet estimateur de λ semble-t-il intéressant ?

1) $(X_1, \ldots, X_n)$ est un n-échantillon de variables indépendantes qui suivent la même loi, et T_n est une fonction de $(X_1, \ldots, X_n)$. Donc T_n est un estimateur de λ.

2) $T_n = \dfrac{1}{n}(X_1 + \ldots + X_n)$ Par linéarité de l'espérance $E(T_n) = \dfrac{1}{n}(E(X_1) + \ldots + E(X_n)) = \dfrac{1}{n} \times n\,\lambda = \lambda$.

$V(T_n) = \dfrac{1}{n^2} V(X_1 + \ldots + X_n) = \dfrac{1}{n^2}\left(V(X_1) + \ldots + V(X_n)\right)$ ($X_1, \ldots, X_n$ indépendantes)

$= \dfrac{1}{n^2} \times n\,\lambda = \dfrac{\lambda}{n}$. $\lim\limits_{n \to +\infty} V(T_n) = \lim\limits_{n \to +\infty} \dfrac{\lambda}{n} = 0$.

L'espérance est égale à θ et la variance tend vers 0 : c'est un bon estimateur.

Rappel : **Loi faible des grands nombres**
Si $(X_i)_{i \geq 1}$ est une suite de variables indépendantes de même loi, qui admettent une espérance m et une variance, et si $T_n = \dfrac{X_1 + \ldots + X_n}{n}$, alors $\forall\ \varepsilon > 0,\ \lim\limits_{n \to +\infty} P\left(\left|T_n - m\right| > \varepsilon\right) = 0$.
En d'autres termes, T_n "tend vers" m. (On dit que (T_n) converge en probabilité vers m)

En Python :
Rappel : Si x est un vecteur qui contient les résultats de n simulations de la variable aléatoire, **np.mean(x)** calcule la moyenne de ces n valeurs.
Si n est grand, la valeur trouvée sera donc très proche de m.

2. Intervalles de confiance

Définition :

Soit $\theta \in \mathbb{R}$ et $\alpha \in]0;1[$. Soit (U_n et V_n) deux estimateurs issus d'un échantillon (X_1, …, X_n).
On dit que $[U_n, V_n]$ est un **intervalle de confiance de θ au niveau de confiance $1 - \alpha$**, (ou au niveau de risque α) si : **$P(U_n \leq \theta \leq V_n) \geq 1 - \alpha$**

Remarques :
_ si T_n est un estimateur de θ et $\varepsilon > 0$,

$$\begin{aligned}
\theta \in [T_n - \varepsilon; T_n + \varepsilon] &\Leftrightarrow T_n - \varepsilon \leq \theta \leq T_n + \varepsilon \\
&\Leftrightarrow -\varepsilon \leq \theta - T_n \leq \varepsilon \\
&\Leftrightarrow \left|\theta - T_n\right| \leq \varepsilon \\
&\Leftrightarrow \left|T_n - \theta\right| \leq \varepsilon \\
&\Leftrightarrow -\varepsilon \leq T_n - \theta \leq \varepsilon \\
&\Leftrightarrow \theta - \varepsilon \leq T_n \leq \theta + \varepsilon
\end{aligned}$$

Pour déterminer un intervalle de confiance, on peut utiliser par exemple **l'inégalité de Bienaymé-Tchebychev**. (Exemple page suivante)

Exemple :
On lance indéfiniment une pièce truquée dont la probabilité de faire pile est $p \in]0;1[$.
Pour tout $i \geq 1$, on considère la VAR X_i définie par : $X_i = 1$ si on obtient pile au i-ème lancer, $X_i = 0$ sinon.
Posons $Z_n = X_1 + \ldots + X_n$ et $Y_n = \dfrac{Z_n}{n} = \dfrac{X_1 + \ldots + X_n}{n}$

On notera $\sigma^2 = p(1-p)$.
1) Montrer que : $\forall\, x \in [0;1]$, $x(1-x) \leq 1/4$
2) Déterminer l'espérance et la variance de Y_n.
3) Montrer que $\forall\, \varepsilon > 0$, $P\left(\left|Y_n - E(Y_n)\right| \leq \varepsilon\right) \geq 1 - \dfrac{1}{4n\varepsilon^2}$.
4) Déterminer un rang n pour lequel $[Y_n - 0{,}01; Y_n + 0{,}01]$ est un intervalle de confiance de p au niveau de confiance 0,95.

1) Soit f la fonction définie sur [0;1] par : $f(x) = x(1-x) = x - x^2$
f est dérivable et $\forall\, x \in [0;1]$, $f'(x) = 1 - 2x$

x	0		1/2		1
f '(x)		+	0	–	
f(x)		↗	1/4	↘	

Donc $\forall\, x \in [0;1]$, $f(x) \leq \dfrac{1}{4}$.

2) $Z_n \longrightarrow \mathcal{B}(n,p)$ (n var de Bernoulli indépendantes)
donc $E(Z_n) = np$ $V(Z_n) = np(1-p) = n\sigma^2$ $\sigma(Z_n) = \sigma\sqrt{n}$
$Y_n = \dfrac{1}{n} Z_n$ donc $E(Y_n) = \dfrac{1}{n} E(Z_n) = \dfrac{np}{n} = p$ $V(Y_n) = \dfrac{1}{n^2} V(Z_n) = \dfrac{\sigma^2}{n}$

3) Par l'inégalité de Bienaymé – Tchebychev : $\forall\, \varepsilon > 0$, $P\left(\left|Y_n - E(Y_n)\right| \leq \varepsilon\right) \geq 1 - \dfrac{V(Y_n)}{\varepsilon^2}$

$P\left(\left|Y_n - p\right| \leq \varepsilon\right) \geq 1 - \dfrac{\sigma^2}{n\varepsilon^2}$ Or $\sigma^2 = p(1-p) \leq \dfrac{1}{4}$ $-\dfrac{\sigma^2}{n\varepsilon^2} \geq -\dfrac{1}{4n\varepsilon^2}$

$P\left(\left|Y_n - E(Y_n)\right| \leq \varepsilon\right) \geq 1 - \dfrac{1}{4n\varepsilon^2}$.

4) $p \in [Y_n - 0{,}01;\, Y_n + 0{,}01] \Leftrightarrow Y_n - 0{,}01 \leq p \leq Y_n + 0{,}01 \Leftrightarrow -0{,}01 \leq p - Y_n \leq 0{,}01$
$\Leftrightarrow \left|p - Y_n\right| \leq 0{,}01 \Leftrightarrow \left|Y_n - p\right| \leq 0{,}01$.

Or $P\left(\left|Y_n - p\right| \leq 0{,}01\right) \geq 1 - \dfrac{100^2}{4n}$

Pour que $P\left(\left|Y_n - p\right| \leq 0{,}01\right) \geq 0{,}95$, il suffit que $1 - \dfrac{100^2}{4n} \geq 0{,}95$

$-\dfrac{100^2}{4n} \geq -0{,}05$ $\dfrac{100^2}{4n} \leq \dfrac{1}{20}$ $4n \geq 100^2 \times 20$ $n \geq 5 \times 100^2$ $n \geq 50\,000$

Définition :

Soit $\theta \in \mathbb{R}$ et $\alpha \in]0;1[$. Soit (U_n et V_n) deux estimateurs issus d'un échantillon (X_1, …, X_n).
On dit que [U_n, V_n] est un **intervalle de confiance asymptotique de θ au niveau de confiance $1 - \alpha$**, (ou au niveau de risque α) si $\lim\limits_{n \to +\infty} P(U_n \leq \theta \leq V_n) \geq 1 - \alpha$

Remarques :
_ Pour trouver un intervalle de confiance asymptotique, on utilise en particulier **le théorème de la limite centrée**
_ Pour plus de simplicité, dans certains exercices, on approche directement la loi par la loi normale (pour n grand)

Ex : Suite de l'exemple précédent
1) Montrer que $\forall\ x > 0$, $\lim\limits_{n \to +\infty} P\left(Y_n - \frac{x\sigma}{\sqrt{n}} \leq p \leq Y_n + \frac{x\sigma}{\sqrt{n}}\right) = 2\Phi(x) - 1$.
2) En déduire un intervalle de confiance asymptotique de p à 95%.

3) Application : On lance 10000 fois la pièce, et on obtient 2345 fois Pile.
Déterminer un intervalle de confiance de p à 95%.

1) $Y_n = \frac{X_1 + \ldots + X_n}{n}$ avec (X_1, …, X_n) indépendants, de même loi et qui admettent une espérance et une variance.
Donc d'après le théorème de la limite centrée
$Y_n^* = \frac{Y_n - E(Y_n)}{\sigma(Y_n)} = \frac{Y_n - p}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(Y_n - p)}{\sigma}$ converge en loi vers $Y \longrightarrow N(0,1)$.

Donc $\lim\limits_{n \to +\infty} P\left(Y_n - \frac{x\sigma}{\sqrt{n}} \leq p \leq Y_n + \frac{x\sigma}{\sqrt{n}}\right) = \lim\limits_{n \to +\infty} P\left(-\frac{x\sigma}{\sqrt{n}} \leq p - Y_n \leq \frac{x\sigma}{\sqrt{n}}\right) =$

$\lim\limits_{n \to +\infty} P\left(-\frac{x\sigma}{\sqrt{n}} \leq Y_n - p \leq \frac{x\sigma}{\sqrt{n}}\right) = \lim\limits_{n \to +\infty} (-x \leq Y_n^* \leq x) = P(-x \leq Y \leq x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$
2) $2\Phi(x) - 1 = 0,95 \Leftrightarrow \Phi(x) = 0,975 \Leftrightarrow x = 1,96$ d'après la table.
Donc avec x = 1,96 : $\lim\limits_{n \to +\infty} P\left(Y_n - \frac{1,96\sigma}{\sqrt{n}} \leq p \leq Y_n + \frac{1,96\sigma}{\sqrt{n}}\right) = 0,95$
$I_n = \left[Y_n - \frac{1,96\sigma}{\sqrt{n}}, Y_n + \frac{1,96\sigma}{\sqrt{n}} \right]$ est un intervalle de confiance asymptotique (mais σ dépend de p).
$\sigma^2 = p(1 - p) \leq \frac{1}{4}$ d'après l'exemple précédent $\sigma \leq \frac{1}{2}$
donc $Y_n + \frac{x\sigma}{\sqrt{n}} \leq Y_n + \frac{x}{2\sqrt{n}}$ et $Y_n + \frac{x}{2\sqrt{n}} \leq Y_n - \frac{x\sigma}{\sqrt{n}}$ donc $I_n \subset \left[Y_n - \frac{0,98}{\sqrt{n}}, Y_n + \frac{0,98}{\sqrt{n}} \right]$, qui est un intervalle de confiance asymptotique de p à 95%.
3) On a donc $Y_{10000} = \frac{2345}{10000} = 0,2345$ donc $\left[0,2345 - \frac{0,98}{\sqrt{10000}}, 0,2345 + \frac{0,98}{\sqrt{10000}} \right]$ est un intervalle de confiance de p à 95%.
0,2345 – 0,0098 = 0,2247 0,2345 + 0,0098 = 0,2445 donc $p \in [0,2247 ; 0,2445]$ avec un niveau de risque de 5%.

3. Exemple : maximum de vraisemblance

Aucune théorie n'est à connaitre dans ce paragraphe.

Soit $\theta \in \Theta$ et $(X_1, .., X_n)$ un n-échantillon, tel que $(X_1, \ldots, X_n)$ sont indépendantes entre elles, et suivent la même loi qu'une variable aléatoire X.

Principe :
On observe dans une expérience que X_1 prend la valeur x_1, X_2 prend la valeur x_2, ….X_n prend la valeur x_n.
On cherche la valeur de θ qui rend maximale la probabilité de cet événement :
Pour $\theta \in \Theta$, on pose $L(\theta) = P((X_1 = x_1) \cap \ldots \cap (X_n = x_n))$
$= P(X_1 = x_1) \times \ldots \times P(X_n = x_n)$ par indépendance.
On exprime $L(\theta)$ en fonction de θ, puis on étudie les variations de cette fonction pour trouver son maximum (on s'aidera souvent de son logarithme).

Ex : On suppose que $X_1, \ldots, X_n$ sont des VAR indépendantes qui suivent toutes la loi de poisson $\mathcal{P}(\theta)$. ($\theta > 0$).
Pour $(x_1, \ldots, x_n) \in \mathbb{N}^n$, on pose $L(\theta) = P((X_1 = x_1) \cap \ldots \cap (X_n = x_n))$ et $f(\theta) = \ln(L(\theta))$.
On pose : $s_n = x_1 + \ldots + x_n$ et $t_n = (x_1)! \times \ldots \times (x_n)!$
1) Exprimer $f(\theta)$ en fonction de θ, n, s_n et t_n.
2) Etudier les variations de f.
3) Montrer que L admet un maximum en une valeur θ_0 que l'on précisera.

1) Par indépendance, $L(\theta) = P(X_1 = x_1)\ldots P(X_n = x_n)$

$$= \frac{\theta^{x1} e^{-\theta}}{x_1!} \times \ldots \times \frac{\theta^{xn} e^{-\theta}}{x_n!} = \frac{\theta^{x1 + \ldots + xn} e^{-n\theta}}{x_1! \ldots x_n!} = \frac{\theta^{sn} e^{-n\theta}}{t_n}$$

en posant $s_n = x_1 + \ldots + x_n$ et $t_n = x_1! \ldots x_n!$
Donc $\ln(L(\theta)) = s_n \ln(\theta) - n\theta - \ln(t_n)$.
Donc $f(\theta) = \ln(L(\theta)) = s_n \ln(\theta) - n\theta - \ln(t_n)$.

2) f est dérivable sur $]0; +\infty($ et $\forall \theta > 0$, $f'(\theta) = \frac{s_n}{\theta} - n = \frac{s_n - n\theta}{\theta}$

θ	0		s_n/n		$+\infty$
$f'(\theta)$		+	0	–	
$f(\theta)$		↗		↘	

Donc f admet un maximum en $\theta_0 = s_n/n$.
$\forall\ \theta > 0$, $f(\theta) \le f(\theta_0)$ $\quad \ln(L(\theta)) \le \ln(L(\theta_0))$
Comme exp est croissante, $L(\theta) \le L(\theta_0)$
Donc le maximum de vraisemblance est $\theta_0 = \frac{x_1 + \ldots + x_n}{n}$.