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 : Avec les hypothèses précédentes
$T_n = \frac{X_1 + \ldots + X_n}{n}$, $T_n = \frac{X_{n-1} + X_n}{2}$, $T_n = \mathrm{Max}(X_1, \ldots, X_n)$, $T_n = (X_1 X_2 \ldots X_n)^{1/n}$ (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 $T_n = \frac{X_1 + \ldots + X_n}{n}$.

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

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 $\forall\ \varepsilon > 0$, $\lim\limits_{n \to +\infty} P\Big(\left|T_n - m\right| > \varepsilon\Big) = 0$.

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 $(X_1, \ldots, X_n)$.
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 : $\mathbf{P(U_n \leq \theta \leq V_n) \geq 1 - \alpha}$

Remarques :

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

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

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.

Définition :

Soit $\theta \in \mathbb{R}$ et $\alpha \in]0;1[$. Soit (U_n et V_n) deux estimateurs issus d'un échantillon $(X_1, \ldots, X_n)$.
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 $\boldsymbol{\lim\limits_{n \to +\infty} P(U_n \leq \theta \leq V_n) \geq 1 - \alpha}$

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%.

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 $(X_1, \dots, X_n)$ 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 $L(\theta) = P((X_1 = x_1) \cap \dots \cap (X_n = x_n))$
$= P(X_1 = x_1) \times \dots \times P(X_n = x_n)$ 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 $X_1, \dots, X_n$ sont des VAR indépendantes qui suivent toutes la loi de poisson $\mathcal{P}(\theta)$. ($\theta >$ 0).
Pour $(x_1, \dots, x_n) \in \mathbb{N}^n$, on pose $L(\theta) = P((X_1 = x_1) \cap \dots \cap (X_n = x_n))$ et $f(\theta) = \ln(L(\theta))$.
On pose : $s_n = x_1 + \dots + x_n$ et $t_n = (x_1)! \times \dots \times (x_n)!$
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.