Leçon 413 : Commentaires

1) Cadre des statistiques, échantillonage

Un phénomène qui dépend du hasard peut être modélisé par une variable aléatoire $X : \Omega \to \mathbb{R}$. D'un point de vue concret, 2 situations peuvent se produire :

- ☞ **La loi de X est connue de l'expérimentateur** : dans ce cas toutes les questions relèvent des probabilités.
- ☞ **La loi de X n'est pas (entièrement) connue de l'expérimentateur, qui dispose simplement d'une suite d'observations** : c'est ici que commencent les statistiques. On cherche alors à construire un modèle probabiliste

Définition 1 : Échantillonage d'une population

Un **échantillon aléatoire** de taille n est un n-uplet $(X_1, \dots, X_n)$ constitué de variables aléatoires réelles **indépendantes et identiquement distribuées**. Leur loi commune est appelée **loi mère** de l'échantillon.

Une **réalisation** de cet échantillon est un n-uplet $(x_1, \dots, x_n) \in \mathbb{R}^n$ tel que $\forall \omega \in \Omega, (X_1(\omega), \dots, X_n(\omega)) = (x_1, \dots, x_n)$.

2) Estimation ponctuelle

On notera généralement :

- ☞ θ le paramètre que l'on souhaite estimer (par exemple la probabilité p d'obtenir *pile* en lançant une pièce).
- ☞ Θ l'ensemble des valeurs possibles pour le paramètre θ.
- ☞ $\mathbf{x} = (x_1, \dots, x_n)$ un échantillon d'observations de la variable X_θ représentant le phénomène que l'on veut modéliser.

Principe :

- ☞ On détermine un réel particulier $\widehat{\theta}$ dont on espère qu'il fournisse une bonne approximation de θ.
- ☞ On cherche ensuite à contrôler l'erreur $|\widehat{\theta} - \theta|$ commise, en général avec un intervalle de confiance.

a. Maximum de vraisemblance (MV)

Définition 2 : Fonction de vraisemblance

La **fonction de vraisemblance** de l'échantillon d'observations $\mathbf{x}$ est l'application :

$$L_{\mathbf{x}} : \begin{array}{ccc} \Theta & \to & \mathbb{R}_+ \\ \theta & \mapsto & \begin{cases} \prod_{i=1}^{n} \mathbb{P}(X_\theta = x_i) & \text{si } X_\theta \text{ est discrète,} \\ \prod_{i=1}^{n} f_{X_\theta}(x_i) & \text{si } X_\theta \text{ est admet une densité } f_{X_\theta} \end{cases} \end{array}$$

Définition 3 : Maximum de vraisemblance pour $\mathbf{x}$

$\widehat{\theta}_{MV} \in \Theta$ est appelé **estimateur de θ par maximum de vraisemblance** si $L_{\mathbf{x}}(\widehat{\theta}_{MV}) = \sup\{L_{\mathbf{x}}(\theta) \mid \theta \in \Theta\}$.

On cherche donc à maximiser la fonction $L_{\mathbf{x}}$. Dans les faits il est souvent plus pratique de maximiser la fonction $H_{\mathbf{x}} = \ln \circ L_{\mathbf{x}}$.

Exercice 1 : Maximum de vraisemblance pour le paramètre p d'une loi (N, p)

On calcule explicitement la fonction $H_{\mathbf{x}}$ définie ci-dessus ainsi que ses deux premières dérivées. Elle est concave donc il suffit de trouver un point critique. Au final $\widehat{p}_{MV} = \dfrac{\overline{\mathbf{x}}_n}{N}$ ce qui est cohérent avec l'intuition donnée par la loi des grands nombres.

b. Théorie des estimateurs ponctuels

Définition 4 : Estimateur sans biais. Soit X_θ une v.a.r. dépendant d'un paramètre θ.

T_n est une **statistique** de X_θ s'il existe un échantillon aléatoire $(X_1, \dots, X_n)$ de X_θ et une fonction $f : \mathbb{R}^n \to \mathbb{R}$ tels que $T_n = f(X_1, \dots, X_n)$.

On dit que T_n est un **estimateur sans biais** de θ si T_n admet une espérance et que $\boxed{\mathbb{E}[T_n] = \theta}$.

On dit que T_n est un **estimateur asymptotiquement sans biais** de θ s'il est le terme général d'une suite $(T_n)_{n\in\mathbb{N}}$ telle que chaque T_n admet une espérance et que $\boxed{\mathbb{E}[T_n] \xrightarrow[n\to+\infty]{} \theta}$.

Définition 5 : Risque quadratique moyen. Soit T_n un estimateur de θ.

On appelle **risque quadratique moyen (RQM)** de T_n, s'il existe, le réel $\boxed{r_\theta(T_n) = \mathbb{E}\left[(T_n - \theta)^2\right]}$.

En particulier si T_n est sans biais, on a alors $\boxed{r_\theta(T_n) = Var(T_n)}$

Propriété 6 : Comparaison de deux estimateurs. Soient S_n et T_n deux estimateurs de θ.

On dit que T_n est un estimateur **plus efficace** que S_n si son RQM est plus faible, i.e. si $r_\theta(T_n) \leqslant r_\theta(S_n)$

☞ **Exemple(s) :**

Soit X une v.a.r. d'espérance μ et $(X_1, \dots, X_n)$ un échantillon aléatoire de X.

La **moyenne empirique** $\overline{X}_n$ est un estimateur sans biais de μ :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

☞ **Exemple(s) :**

Soit X une v.a.r. de variance σ^2 et $(X_1, \dots, X_n)$ un échantillon aléatoire de X.

La **variance empirique corrigée** S_n^2 est un estimateur sans biais de σ^2 :

$$S_n^2 = \frac{1}{\mathbf{n-1}}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Définition 7 : Estimateur correct

On dit qu'une suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ de θ est **correcte** si elle est sans biais et convergente, c'est-à-dire si :

$$T_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$$

Propriété 8 :

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de θ telle que $r_\theta(T_n) \xrightarrow[n\to+\infty]{} 0$. Alors $(T_n)_{n\in\mathbb{N}^*}$ est correcte.

Exercice 2 : Estimation du nombre de Panzers

Les X_k suivent une loi $\mathcal{U}(\llbracket 1; N \rrbracket)$ et on calcule 2 estimateurs sans biais :

- ☞ $T_n = 2\overline{X}_n - 1$ qui utilise la moyenne empirique des numéros des chars capturés ;
- ☞ M_n qui prend tout simplement le max des numéros.

L'exercice n'est pas très difficile mais fait appel à de nombreuses notions et pas mal de calculs de probabilités. Dans la dernière question on aborde la notion de « réalisations » d'un estimateur et si on note t_5 et m_5 les réalisations respectives de T_5 et M_5 on trouve $t_5 = 69$ et $m_5 = 64$. Ce n'est évidemment pas un exemple réaliste pour des raisons de simplicité de calcul.

3) Estimation par intervalles de confiance

Théorème 9 : Théorème central limite

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a.r. indépendantes et identiquement distribuées admettant un moment d'ordre 2. On note μ leur espérance et σ^2 leur variance. On a alors :

$$Y_n = \frac{\sum_{i=1}^{n} Xi - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

En particulier on a alors :

$$\forall \gamma \in [0,1[, \exists! z_\gamma \in \mathbb{R}+ \text{ tel que } \mathbb{P}\left(-z_\gamma \leqslant Y_n \leqslant z_\gamma\right) = \gamma$$

Et en pratique on utilise souvent $z_{0,95} \approx 1,96$.

Exercice 3 : Approximation de π par des méthodes de Monte-Carlo (DÉV)

On crée dans cet exercice 2 estimateurs sans biais de $\frac{\pi}{4}$ à l'aide des méthodes de Monte-Carlo et on les compare. Dans la dernière question, on utilise la notion d'intervalle de confiance pour déterminer la taille de la simulation nécessaire à obtenir un résultat satisfaisant. Plus de détails seront donnés dans le développement.

Exercice 4 : Fiabilité d'un vaccin

On considère ici un échantillon aléatoire $(X_1,\dots,X_n)$ tel que $X_i = 1$ si la i-ème personne n'est pas immunisée et 0 sinon, donc de loi $\mathcal{B}(p)$, avec p le paramètre inconnu que l'on souhaite estimer. À l'aide du TCL, on écrit l'intervalle de confiance que l'on peut majorer à l'aide de l'énoncé. C'est assez similaire à ce qui est fait dans la dernière question du développement, mais de façon plus générale.

4) Tests paramétriques

On émet une **hypothèse nulle** (H_0) sur une v.a.r. X et on dispose de sa négation (H_1), l'hypothèse alternative. On se fixe un risque $\alpha \in]0;1[$ et on observe une réalisation $\mathbf{x} = (x_1,\dots,x_n)$ d'un échantillon de taille n de X. Selon le résultat, deux possibilités :

- On rejette (H_0) avec un risque α de se tromper ;
- On a pas de raison de rejeter (H_0) au risque α.

On peut utiliser un intervalle de confiance pour tester une hypothèse.

Exercice 5 : Test paramétrique pour une loi de Bernoulli

Ici on teste l'hypothèse (H_0) : « la molécule est efficace ». Cela revient à estimer le paramètre p d'une loi $\mathcal{B}(p)$ et on a alors $(H_0) : p \geqslant 0,99$.