CHAPITRE 14

Estimation statistique

14.1 Estimation ponctuelle

14.1.1 Introduction

Exemples :

E1 – On cherche à calculer la taille moyenne d'un homme de 30 ans en France. Cependant, il est impossible de déterminer exactement la taille de tous les hommes français et d'en faire une moyenne. Pour donner une valeur approchée de cette moyenne, on va prendre un échantillon d'hommes, par exemple 1000 hommes, on détermine leur taille puis on fait la moyenne.
Avec un échantillon assez grand, on considère que l'on a obtenu une valeur approchée, c'est-à-dire une **estimation**, de la taille moyenne d'un homme de 30 ans.

E2 – On dispose d'une urne qui contient des boules rouges et des boules blanches, mais on ne connaît pas la composition de l'urne. On procède à 100 tirages avec remise, et on obtient 30 boules rouges et 70 boules blanches. A combien peut-on estimer la proportion de boules rouges dans l'urne ?

Définition 1
On considère une espérience aléatoire E et une variable aléatoire réelle X qui lui est liée. On ne connaît pas la loi de X, mais on sait qu'elle appartient à une famille de lois dépendant d'un paramètre θ réel, qui appartient à un ensemble $\triangle$ (par exemple, X suit une loi de Poisson de paramètre λ inconnu, mais on sait que $\lambda \in \mathbb{R}^+$).
Lorsqu'on réalise une fois l'expérience E, la valeur que prend X, souvent notée x, s'appelle **une réalisation de X**.
Le but ici est de donner une valeur approchée de θ à l'aide de la donnée de n réalisations de X, que l'on notera $(x_1, \ldots, x_n)$.
Le n-uplet $(x_1, \ldots, x_n)$ s'appelle un **n-échantillon de données**.

Exemple :

On reprend l'exemple précédent avec les boules rouges/blanches.

On choisit ici X la variable qui vaut 1 si on tire une boule rouge et 0 sinon. Alors, on sait que X suit une loi de Bernoulli de paramètre p, où p est la probabilité de tirer une boule rouge, c'est-à-dire que p représente la proportion de boules rouges. Dans cette expérience, on cherche à connaître la valeur de p. On a donc ici $\theta = p$ et $\Delta =]0, 1[$.
L'énoncé de l'exemple précédent nous dit que lors de 100 réalisations de la variable X, X a pris 30 fois la valeur 1 et 70 fois la valeur 0.

Remarque :

Il y a deux types d'estimations que nous allons voir :
- soit chercher une valeur approchée de θ
- soit cherche un intervalle dans lequel θ a une forte probabilité de se trouver

14.1.2 Echantillon

Définition 2

Soit X une variable aléatoire réelle, définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.
On appelle **n-échantillon** de la variable X tout n-uplet $(X_1, \dots, X_n)$ de variables aléatoires indépendantes définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ et de même loi que X.

Remarque :

À chaque fois que l'on réalise n fois l'expérience E, on obtient des échantillons de données différents. On peut donc définir les variables aléatoires $X_1, \dots, X_n$ définies de la façon suivante : à chaque réalisation de n fois l'expérience E, X_i correspond à la valeur prise par X lors de la i-ième réalisation de E.

14.1.3 Estimateurs

Définition 3

Soit $(X_1, \dots, X_n)$ un n-échantillon d'une variable X dont la loi dépend d'un paramètre θ.
On appelle **estimateur de θ** toute variable aléatoire T_n qui est une fonction des variables $X_1, \dots, X_n$

$$T_n = f(X_1, \dots, X_n)$$

Lorsqu'on a un échantillon de données $(x_1, \dots, x_n)$, $f(x_1, \dots, x_n)$ est une **estimation de θ**.

Définition 4

Soit $(X_1, \dots, X_n)$ un n-échantillon d'une variable X dont la loi dépend d'un paramètre θ.
La variable $\overline{X_n}$ définie par

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

est un estimateur de θ. Cette variable est appelée la **moyenne empirique** de l'échantillon.

Exemple :

Reprenons notre exemple fil-rouge. Un échantillon de données est une liste de 0 et de 1, où on a 30 fois le nombre 1 et 70 fois le nombre 0. Donc, après nos 100 tirages, la moyenne empitique de X vaut

$$\frac{30 \times 1 + 70 \times 0}{100} = 0.3$$

Une estimation de p est donc 0.3.

14.1.4 Biais d'un estimateur

Définition 5

Soit T_n un estimateur de θ. Si T_n admet une espérance pour tout θ, alors on appelle **biais de l'estimateur**, le réel

$$b(T_n) = \mathbb{E}(T_n - \theta) = \mathbb{E}(T_n) - \theta$$

Lorsque $b(T_n) = 0$, autrement-dit lorsque $\mathbb{E}(T_n) = \theta$, on dit que T_n est un **estimateur sans biais**.

Remarques :

- R1 – Le biais mesure l'écart moyen entre les valeurs prises par l'estimateur et le réel que l'on cherche à estimer.
- R2 – Lorsqu'on dit que l'estimateur est sans biais, cela signifie qu'en moyenne, les valeurs de l'estimateur sont très proches de θ.
- R3 – Rien n'empêche un estimateur sans biais de prendre des valeurs très éloignées de θ, car en moyenne les écarts peuvent se compenser.

Exemple :

On lance une pièce et on note $X = 1$ si on obtient Pile, et $X = 0$ sinon. Alors X suit une loi de Bernoulli de paramètre $p = \mathbb{P}(\text{Pile})$. On se donne un n-échantillon $(X_1, \dots, X_n)$ de la variable X et on considère l'estimateur $T_n = X_1$.
On a bien $\mathbb{E}(T_n) = \mathbb{E}(X_1) = \mathbb{E}(X) = p$, donc T_n est un estimateur sans biais de p.
Une estimation de p est donc une valeur prise par T_n, mais pourtant les valeurs possibles de T_n sont 0 et 1, qui sont bien éloignées de p...

Remarque :

On voit bien avec l'exemple précédent qu'avoir un estimateur est parfois trop léger. Il faut pouvoir classer les différentes estimateurs pour savoir lequel est le "meilleur".

14.1.5 Risque quadratique

Définition 6

Soit T_n un estimateur de θ. Si T_n admet un moment d'ordre 2 pour tout θ, alors on appelle **risque quadratique de l'estimateur** T_n le réel

$$r(T_n) = \mathbb{E}\left((T_n - \theta)^2\right)$$

Remarque :

Le risque quadratique mesure la moyenne de l'écart de T_n à θ au carré. Comme un carré est toujours positif, les écarts à θ "en plus" ou "en moins" ne peuvent plus se compenser, mais se cumulent.
On a donc bien ici une façon de mesurer si T_n est un "bon" estimateur de θ.

Théorème 7

Soit T_n un estimateur de θ admettant un moment d'ordre 2. Alors,

$$r(T_n) = \mathbb{V}(T_n) + (b(T_n))^2$$

Démonstration :

$r(T_n) = \mathbb{E}\left((T_n - \theta)^2\right) = \mathbb{E}\left(T_n^2 - 2\theta T_n + \theta^2\right) = \mathbb{E}(T_n^2) - 2\theta\mathbb{E}(T_n) + \theta^2$
$= \mathbb{V}(T_n) + (\mathbb{E}(T_n))^2 - 2\theta\mathbb{E}(T_n) + \theta^2 = \mathbb{V}(T_n) + (\mathbb{E}(T_n) - \theta)^2 = \mathbb{V}(T_n) + (b(T_n))^2$

Remarque :

Si T_n est un estimateur sans biais, alors $r(T_n) = \mathbb{V}(T_n)$.

Exemple :

On revient à notre exemple Pile/Face avec $p = \mathbb{P}(\text{Pile})$ et X qui suit une loi de Bernoulli de paramètre p. Soit $(X_1, \dots, X_n)$ un n-échantillon de X.
Si on considère l'estimateur $T_n = X_1$, le risque quadratique est

$$r(T_n) = \mathbb{V}(T_n) = \mathbb{V}(X_1) = pq$$

On considère maintenant la moyenne empirique $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$.
On a $E(\overline{X_n}) = \dfrac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \dfrac{1}{n} \times np = p$, donc $b(\overline{X_n}) = 0$. De plus, comme les X_i sont indépendantes,

$$r(\overline{X_n}) = \mathbb{V}(\overline{X_n}) = \frac{1}{n^2} n\mathbb{V}(X_1) = \frac{pq}{n}$$

Le risque quadratique de $\overline{X_n}$ est n-fois plus petit que celui de T_n et de plus, plus n est grand, plus $r(\overline{X_n})$ est petit. Ainsi, $\overline{X_n}$ est un bien meilleur estimateur que T_n.

14.1.6 Estimation de l'espérance d'une VAR

Théorème 8

Soit X une variable aléatoire admettant une espérance m inconnue et admettant une variance. Soit $(X_1, \dots, X_n)$ un n-échantillon de X. Alors, la moyenne empirique

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

est un estimateur sans biais de m, et dont le risque quadratique tend vers 0 lorsque n tend vers $+\infty$.

Démonstration :

$$\mathbb{E}(\overline{X_n}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n} \times nm = m$$

donc $b(\overline{X_n}) = 0$. Comme les X_i sont indépendantes, on a

$$r(\overline{X_n}) = \mathbb{V}(\overline{X_n}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2} n\mathbb{V}(X) = \frac{\mathbb{V}(X)}{n}$$

Donc $r(\overline{X_n}) \underset{n\to+\infty}{\longrightarrow} 0$.

Remarque :

La moyenne empirique $\overline{X_n}$ est donc un très bon estimateur de l'espérance, et il est de plus en plus meilleur lorsque l'échantillon devient grand.

14.2 Intervalles de confiance

14.2.1 Définition

Définition 9
Soit $(X_1, \dots, X_n)$ un n-échantillon d'une loi μ_θ et U_n et V_n deux estimateurs de θ. Pour tout réel $\alpha \in]0,1[$, on dit que $[U_n, V_n]$ est un **intervalle de confiance de θ au niveau de confiance $1-\alpha$** (ou au **risque α**), si on a

$$\mathbb{P}(U_n \leqslant \theta \leqslant V_n) \geqslant 1-\alpha$$

14.2.2 Utilisation de Bienaymé-Tchebychev

Proposition 10
Soit T_n un estimateur sans biais de θ, admettant un moment d'ordre 2.
On suppose qu'il existe $v_n \in \mathbb{R}$ (indépendant de θ) tel que $\mathbb{V}(T_n) \leqslant v_n$. Alors pour tout $\varepsilon \in \mathbb{R}^+$, l'intervalle

$$[T_n - \varepsilon, T_n + \varepsilon]$$

est un intervalle de confiance de θ au niveau de confiance $1 - \dfrac{v_n}{\varepsilon^2}$.

Démonstration :

En effet, on utilise l'inégalité de Bienaymé-Tchebychev : $\mathbb{P}(|T_n - \theta| \leqslant \varepsilon) \geqslant 1 - \dfrac{\mathbb{V}(T_n)}{\varepsilon^2}$,
autrement dit $\mathbb{P}(T_n - \varepsilon \leqslant \theta \leqslant T_n + \varepsilon) \geqslant 1 - \dfrac{\mathbb{V}(T_n)}{\varepsilon^2} \geqslant 1 - \dfrac{v_n}{\varepsilon^2}$.

Exemples :

E 1 – On considère un n-échantillon $(X_1, \dots, X_n)$ d'une loi de Bernoulli de paramètre θ et $\overline{X_n}$ la moyenne empirique associée à l'échantillon. On a alors

$$\mathbb{V}(\overline{X_n}) = \frac{\theta(1-\theta)}{n} \leqslant \frac{1}{4n}$$

et donc

$$\mathbb{P}(\overline{X_n} - \varepsilon \leqslant \theta \leqslant \overline{X_n} + \varepsilon) \geqslant 1 - \frac{1}{4n\varepsilon^2}$$

Pour obtenir un intervalle de confiance au niveau de confiance 0.95, il faut prendre ε tel que

$$\frac{1}{4n\varepsilon^2} = 0.05$$

autrement dit

$$\varepsilon = \frac{1}{\sqrt{0.2n}}$$

Donc $\left[\overline{X_n} - \dfrac{1}{\sqrt{0.2n}}, \overline{X_n} + \dfrac{1}{\sqrt{0.2n}}\right]$ est un intervalle de confiance de θ au niveau de confiance 0.95.

E 2 – Revenons à notre exemple fil-rouge sur les tirages dans l'urne. Une réalisation de $\overline{X_n}$ était $\dfrac{30}{100}$ et on avait $n = 100$, donc un intervalle de confiance réalisé pour p est approximativement $[0.076, 0.524]$.

14.2.3 Utilisation de la loi normale

Proposition 11

On considère un n-échantillon $(X_1, \dots, X_n)$ d'une loi d'espérance m (inconnue) et de variance σ^2 (connue). Alors, un intervalle de confiance de m au niveau de confiance $1-\alpha$ est

$$\left[\overline{X_n} - \frac{\sigma t_\alpha}{\sqrt{n}}, \overline{X_n} + \frac{\sigma t_\alpha}{\sqrt{n}}\right]$$

où $\overline{X_n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ *et où* t_α *est le plus petit* $t > 0$ *tel que*

$$\mathbb{P}(-t \leqslant \overline{X_n}^* \leqslant t) \geqslant 1 - \alpha$$

avec $\overline{X_n}^*$ *la variable centrée réduite associée à* X_n.

Démonstration :

On a

$$\mathbb{E}(\overline{X_n}) = m \qquad \text{et} \qquad \mathbb{V}(\overline{X_n}) = \frac{\sigma^2}{n}$$

D'après le Théorème de la Limite Centrée, on sait que la variable

$$\overline{X_n^*} = \frac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}}$$

converge en loi vers la loi normale centrée réduite $\mathcal{N}(0,1)$. Donc pour n assez grand ($n \geqslant 30$), on peut assimiler la loi de $\overline{X_n}^*$ à la loi normale centrée réduite.

On cherche donc une valeur approchée du plus petit t_α tel que $\mathbb{P}(-t_\alpha \leqslant \overline{X_n}^* \leqslant t_\alpha) \geqslant 1-\alpha$, autrement dit

$$\Phi(t_\alpha) - \Phi(-t_\alpha) = 2\Phi(t-\alpha) - 1 \geqslant 1 - \alpha$$

donc

$$\Phi(t_\alpha) \geqslant 1 - \frac{\alpha}{2}$$

On a donc trouvé t_α tel que

$$\mathbb{P}\left(t_\alpha \leqslant \frac{\overline{X_n} - m}{\sigma/\sqrt{n}} \leqslant t_\alpha\right) \geqslant 1 - \alpha$$

autrement dit

$$\mathbb{P}\left(\overline{X_n} - \frac{\sigma t_\alpha}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + \frac{\sigma t_\alpha}{\sqrt{n}}\right) \geqslant 1 - \alpha$$

Remarque :

Par exemple, pour $\alpha = 0.05$, on a $1 - \dfrac{\alpha}{2} = 0.975$, donc $t_\alpha \simeq 1.96$

Exemple :

Dans notre exemple fil-rouge, une réalisation de $\overline{X_n}$ est 0.3 et on a $n = 100$. Si on suppose de plus que l'on a un écart-type maximal, c'est-à-dire égal à 0.5, alors un intervalle de confiance réalisé pour p au niveau de confiance 0.95 est donc $[0.202\,,\, 0.398]$, donc l'intervalle est meilleur que celui obtenu avec Bienaymé-Tchebychev.