Estimations

I Introduction

Exemple 1:

Si on cherche à calculer la taille moyenne d'un homme de 40 ans en France, 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 prend 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, « une estimation », de la taille moyenne d'un homme de 40 ans.

Exemple 2:

On dispose d'une urne contenant des boules rouges et des boules blanches mais on ne connait pas la composition de l'urne. En 100 tirages avec remise on obtient 30 boules rouges et 70 boules blanches. A combien peut-on estimer la proportion de boules rouges dans l'urne ?

Mise en place de la théorie :

On considère une expérience aléatoire e et une variable aléatoire réelle X qui lui est liée. On ne connait pas la loi de X mais on sait qu'elle appartient à une famille de loi dépendant d'un paramètre θ réel qui appartient à un ensemble Θ. (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 $(x_1, \cdots, x_n)$. Le n-uplet $(x_1, \cdots, x_n)$ s'appelle un **n-échantillon de données**.

Retour à l'exemple 2

Prenons 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 est la proportion de boules rouges. Dans cette expérience on cherche à connaitre la valeur de p. On a donc ici $\theta = p$ et $\Theta =]0;1[$.

L'énoncé de l'exemple 2 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.

Nous allons voir deux types d'estimations : soit on cherchera une valeur approchée de θ soit nous chercherons un intervalle dans lequel θ a une forte probabilité de se trouver.

II Estimation ponctuelle

1 n-échantillon

Définition 1

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

La loi de X est appelée la **loi parente** de l'échantillon

A 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 $X_1, \cdots, X_n$ 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-ème réalisation de e.

2 Estimateur-Estimation

Définition 2
Soit $(X_1, \cdots, X_n)$ 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 $(X_1, \cdots, X_n)$:

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

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

Définition 3
Soit $(X_1, \cdots, X_n)$ un n-échantillon d'une variable X dont la loi dépend d'un paramètre θ.
La variable $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$ est un estimateur de θ. Cette variable est souvent appelée **moyenne empirique** de l'échantillon.

Exemple 3:
Si on revient à notre exemple 2, un échantillon de donnée 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 empirique de X vaut : $\dfrac{30 \times 1 + 70 \times 0}{100} = 30\%$.
Une estimation de p est donc 0,3.

3 Qualité de l'estimateur

a Biais

Définition 4
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

$$b(T_n) = E(T_n - \theta) = E(T_n) - \theta$$

Lorsque $b(T_n) = 0$, ce qui équivaut à $E(T_n) = \theta$, on dit que T_n est **un estimateur sans biais**.

Remarques :
- Le biais mesure l'écart moyen entre les valeurs prises par l'estimateur et le réel que l'on cherche à estimer.
- Lorsqu'on dit que l'estimateur est sans biais cela signifie que en moyenne les valeurs de l'estimateur sont très proches de θ.
- Attention rien n'empêche un estimateur sans biais de prendre des valeurs très éloignés de θ car en moyenne les écarts peuvent se compenser.
  Exemple 4:
  On lance une pièce et on note $X = 1$ si on obtient pile et $X = 0$ sinon. X suit une loi de Bernoulli de paramètre $p = P(pile)$. On se donne un n-échantillon $(X_1, \cdots, X_n)$ de la variable X et on considère l'estimateur $T_n = X_1$.
  On a bien $E(T_n) = E(X_1) = E(X) = p$ 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...

  On a donc besoin d'une mesure supplémentaire pour différencier deux estimateurs sans biais et dire lequel est le « meilleur ».

b Risque quadratique

Définition 5
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** le réel :

$$r(T_n) = E([T_n - \theta]^2)$$

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 1
Soit T_n un estimateur de θ admettant un moment d'ordre 2. Alors on a

$$r(T_n) = V(T_n) + (b(T_n))^2$$

Démonstration :

$$\begin{aligned} r(T_n) &= E([T_n - \theta]^2) = E(T_n^2 - 2\theta T_n + \theta^2) \\ &= E(T_n^2) - 2\theta E(T_n) + \theta^2 \quad \text{or } E(T_n^2) = V(T_n) + E(T_n)^2 \\ &= V(T_n) + (E(T_n))^2 - 2\theta E(T_n) + \theta^2 \\ &= V(T_n) + (E(T_n) - \theta)^2 = V(T_n) + (b(T_n))^2 \end{aligned}$$

□

Remarque :
Si T_n est un estimateur sans biais alors $r(T_n) = V(T_n)$.

Exemple 5:
On revient à notre pile ou face avec p la probabilité d'obtenir pile et X qui suit une loi de Bernoulli de paramètre p. Soit $(X_1, \cdots, X_n)$ un n-échantillon de X.

• Si on considère l'estimateur $T_n = X_1$, le risque quadratique est $r(T_n) = V(T_n) = V(X_1) = pq$.

• On considère maintenant la moyenne empirique $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$.

On a $E(\overline{X_n}) = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n} \times np = p$ donc $b(\overline{X_n}) = 0$.

De plus, comme les X_i sont indépendantes,

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

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 $r(\overline{X_n})$ sera petit.

4 Estimation de l'espérance

Propriété 1
On considère X une variable aléatoire d'espérance m et $(X_1, \cdots, X_n)$ un n-échantillon de X. Alors la moyenne empirique $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$ est un estimateur sans biais de m et dont le risque quadratique tend vers 0 lorsque n tend vers $+\infty$.

Démonstration :

On a $E(\overline{X_n}) = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n} \times nm = m$ donc $b(\overline{X_n}) = 0$.

Comme les X_i sont indépendantes $r(\overline{X_n}) = V(\overline{X_n}) = \dfrac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \dfrac{1}{n^2} nV(X) = \dfrac{V(X)}{n}$.

Donc $r(\overline{X_n}) \to 0$ lorsque $n \to +\infty$

□

III Intervalle de confiance

1 *Définition*

Définition 6
Soit $(X_1, \cdots, X_n)$ 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 aussi au risque α) si on a

$$P([U_n \leqslant \theta \leqslant V_n]) \geqslant 1 - \alpha$$

2 *Construction pratique*

a Utilisation de l'inégalité de Bienaymé-Tchebychev

Si on considère T_n un estimateur sans biais de θ alors l'inégalité de Bienamyé-Tchebychev s'écrit :

$$P(|T_n - \theta| \leqslant \varepsilon) \geqslant 1 - \frac{V(T_n)}{\varepsilon^2} \Leftrightarrow P(T_n - \varepsilon \leqslant \theta \leqslant T_n + \varepsilon) \geqslant 1 - \frac{V(T_n)}{\varepsilon^2}$$

$V(T_n)$ dépend de θ mais si on arrive à majorer $V(T_n) \leqslant v_n$ alors on aura

$$P(T_n - \varepsilon \leqslant \theta \leqslant T_n + \varepsilon) \geqslant 1 - \frac{v_n}{\varepsilon^2}$$

Donc $[T_n - \varepsilon; T_n + \varepsilon]$ est un intervalle de confiance de θ au niveau de confiance $1 - \dfrac{v_n}{\varepsilon^2}$.

Exemple 6:

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

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

et donc

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

Pour obtenir un intervalle de confiance au niveau de confiance $0,95$, il faut prendre ε tel que

$$\frac{1}{4n\varepsilon^2} = 0,05 \Leftrightarrow \varepsilon = \frac{1}{\sqrt{0,2n}}$$

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.

• Revenons à notre exemple 2, avec nos tirages, 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 $[0,076; 0,524]$

b Intervalle de confiance pour une moyenne à l'aide de la loi normale

On considère un n-échantillon $(X_1, \cdots, X_n)$ d'une loi d'espérance m (inconnue) de variance σ^2 connue. On pose $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$ et on cherche un intervalle de confiance de m au niveau de confiance $1-\alpha$.

On a $E(\overline{X_n}) = m$ et $V(\overline{X_n}) = \dfrac{\sigma^2}{n}$.

D'après le théorème de la limite centrée, $\overline{X_n}^* = \dfrac{\overline{X_n} - m}{\sigma/\sqrt{n}}$ converge en loi vers la loi normale centrée réduite.

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

Première étape :

On cherche une valeur approchée du plus petit t_α tel que :

$$\begin{aligned} & P(-t_\alpha \leqslant \overline{X_n}^* \leqslant t_\alpha) \geqslant 1-\alpha \\ \Leftrightarrow & \Phi(t_\alpha) - \Phi(-t_\alpha) \geqslant 1-\alpha \\ \Leftrightarrow & 2\Phi(t_\alpha) - 1 \geqslant 1-\alpha \\ \Leftrightarrow & \Phi(t_\alpha) \geqslant 1 - \frac{\alpha}{2} \end{aligned}$$

Quelle est la valeur de t_α pour $\alpha = 0,05$?

On a $1 - \dfrac{\alpha}{2} = 0,975$ *donc* $t_\alpha \approx 1,96$

Deuxième étape :

On a donc trouvé t_α tel que :

$$\begin{aligned} & P(t_\alpha \leqslant \overline{X_n}^* \leqslant t_\alpha) \geqslant 1-\alpha \\ \Leftrightarrow & P(t_\alpha \leqslant \frac{\overline{X_n} - m}{\sigma/\sqrt{n}} \leqslant t_\alpha) \geqslant 1-\alpha \\ \Leftrightarrow & P\left(\overline{X_n} - \frac{\sigma t_\alpha}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + \frac{\sigma t_\alpha}{\sqrt{n}}\right) \geqslant 1-\alpha \end{aligned}$$

Donc $\left[\overline{X_n} - \dfrac{\sigma t_\alpha}{\sqrt{n}}; \overline{X_n} + \dfrac{\sigma t_\alpha}{\sqrt{n}}\right]$ est un intervalle de confiance de m au niveau de confiance $1-\alpha$.

Exemple 7:

Dans notre exemple 2, 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 $[0,202; 0,398]$ (meilleur qu'avec Bienaymé-Tchebychev).