Chapitre 12. Estimation

Table des matières

Dans ce chapitre, on applique les résultats obtenus en probabilités à des études statistiques : On extrait un échantillon d'une certaine population, et on mesure pour chaque individu de cet échantillon la valeur d'une certaine variable.
Quels renseignements, concernant les caractéristiques de cette variable, peut-on déduire de cette mesure ?

Exemple : Soit $\mathcal{P}$ la population française adulte, X la taille d'un Français pris au hasard, μ la taille moyenne des Français.
On veut évaluer ou **estimer**, cette moyenne μ.
Pour cela on prend un échantillon de taille n, qui fournit n valeurs $x_1, x_2, \ldots, x_n$ de la variable X (valeurs *mesurées*). (x_i étant la valeur de X pour l'individu n°i de l'échantillon étudié.)

1 Modèle mathématique

On suppose que la loi de probabilité de la variable aléatoire X associée à l'observation d'un certain phénomène aléatoire appartient à une famille de lois, dépendant d'un paramètre θ réel, $\theta \in \Theta \subset \mathbb{R}$.
Par exemple : on sait que X suit une loi de Poisson, mais on ne connait pas son paramètre λ, on sait seulement que $\lambda \in \mathbb{R}^{+*}$.
On suppose qu'il existe un espace probabilisé $(\Omega, \mathcal{A}, P)$ sur lequel on peut définir une suite de variables aléatoire indépendantes et de même loi que X, notée $(X_n)_{n\in\mathbb{N}^*}$.

Définition 1 :

$(X_1, X_2, \ldots, X_n)$ est un n-échantillon de V.A.R. indépendantes et de même loi que X.

En pratique : X_i est la variable aléatoire de même loi que X liée à l'individu n°i de l'échantillon statistique.

Attention : $X_i \neq x_i$ (X_i est une variable aléatoire, et x_i est une valeur mesurée.)

2 Estimateur

Définition 2 :

On appelle **estimateur** de θ toute suite de V.A.R. $(T_n)_{n\geq 1}$ où $T_n = g(X_1, X_2, \ldots, X_n)$, g étant une fonction de n variables.
On dira, par abus de langage, que T_n est un estimateur de θ.

En pratique : On ne prend évidemment pas n'importe quelle fonction g : l'objectif du choix d'un estimateur est d'avoir une variable aléatoire qui se "rapproche" du paramètre qu'on doit estimer, dans un sens que l'on va préciser.

Définition 3 :

Si T_n est un estimateur de θ, le **biais** de cet estimateur est le nombre réel : $b(T_n) = E(T_n - \theta) = E(T_n) - \theta$.
T_n est un estimateur sans biais de θ si $b(T_n) = 0$ (c'est-à-dire $E(T_n) = \theta$).

Exemple 1 : Soit $X \hookrightarrow \mathcal{B}(p)$, avec un paramètre p inconnu.
Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de X sur une population $\mathcal{P}$.
Alors si $T_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, la variable T_n est un estimateur de p.

Rappel : La loi faible des grands nombres permet de dire que la suite (T_n) converge en probabilité vers la variable certaine de valeur p.
De plus, $E(T_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} E(X_i) = \frac{1}{n}np = p$.
Donc T_n est un estimateur sans biais de p.

Exemple 2 : Soit $X \hookrightarrow \mathcal{P}(\lambda)$, où λ est inconnu.
On pose $T_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, où $(X_1, X_2, \ldots, X_n)$ est un n-échantillon de X.
$E(T_n) = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n} n\lambda = \lambda$.
T_n est donc un estimateur sans biais de λ.

De façon générale :

Théorème 1 : Si X suit une loi de probabilité telle que $E(X) = m$, et si $T_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, alors T_n est un estimateur sans biais de m.

Exemple 3 : Soit X une variable aléatoire d'espérance m, de variance σ^2. Un n-échantillon $(X_1, X_2, \ldots, X_n)$ étant donné, on pose :
$T_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ et $Y_n = \dfrac{1}{n}\sum_{k=1}^{n}(X_k - T_n)^2$
(Y_n est la variance statistique de l'échantillon.)
Y_n est-il un estimateur sans biais de σ^2 ?
Pour le savoir il faut calculer $E(Y_n)$ et comparer avec σ^2.

$$\begin{aligned} E(Y_n) &= \frac{1}{n}\sum_{k=1}^{n} E\left[(X_k - T_n)^2\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} E(X_k^2 - 2T_nX_k + T_n^2) \\ &= \frac{1}{n}\sum_{k=1}^{n} E(X_k^2) - \frac{2}{n}\sum_{k=1}^{n} E(T_nX_k) + E(T_n^2) \end{aligned}$$

$E(X_k^2) = V(X_k) + (E(X_k))^2 = \sigma^2 + m^2$
$E(T_n^2) = V(T_n) + (E(T_n))^2 = \dfrac{1}{n^2}\sum_{k=1}^{n} V(X_k) + m^2 = \dfrac{\sigma^2}{n} + m^2$
$E(X_kT_n) = E\left(\dfrac{1}{n}\sum_{i=1}^{n} X_kX_i\right) = \dfrac{1}{n}\left(E(X_k^2) + \sum_{i\neq k} E(X_k)E(X_i)\right) = \dfrac{1}{n}(\sigma^2 + m^2 + (n-1)m^2) = \dfrac{\sigma^2}{n} + m^2$
d'où : $E(Y_n) = \dfrac{1}{n}[n(\sigma^2 + m^2)] - \dfrac{2}{n}(\sigma^2 + nm^2) + \dfrac{\sigma^2}{n} + m^2 = \sigma^2\left(1 - \dfrac{1}{n}\right) = \dfrac{n-1}{n}\sigma^2$
donc Y_n n'est pas un estimateur sans biais de σ^2.
Par contre, si on pose $Z_n = \dfrac{n}{n-1}Y_n$, alors $E(Z_n) = \dfrac{n}{n-1}E(Y_n) = \sigma^2$
Z_n est un estimateur sans biais de σ^2.

Définition 4 : Soit T_n un estimateur de θ, admettant un moment d'ordre 2 ($E(T_n^2)$ existe).
On appelle risque quadratique de T_n le réel : $r(T_n) = E((T_n - \theta)^2)$

Théorème 2 : $r(T_n) = V(T_n) + [b(T_n)]^2$

dém :

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

Remarque : Trouver un "bon" estimateur de θ revient à trouver un estimateur pour lequel $r(T_n)$ est le plus faible possible.
Pour un estimateur sans biais, $r(T_n) = V(T_n)$: il faut alors minimiser la *variance* de l'estimateur.
Sinon, on peut éventuellement choisir un estimateur de biais non nul, et jouer sur les deux quantités $V(T_n)$ et $b^2(T_n)$.

Exemple 4 : Soit $X \hookrightarrow \mathcal{E}\left(\frac{1}{m}\right)$, où m est inconnu.
On remarque que : $E(X) = m$.
X est la durée de vie moyenne d'une lampe. On cherche à estimer la durée de vie moyenne m. Pour cela, on considère un n-échantillon $(X_1, X_2, \ldots, X_n)$ de n lampes choisies au hasard.
1°) Soit $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Montrer que T_n est un estimateur sans biais de m et calculer son risque quadratique.
2°) On pose $Y = inf(X_1, X_2, \ldots, X_n)$. Déterminer la fonction de répartition F_Y de Y, et en déduire sa loi.
Soit $Z = nY$. Montrer que Z est un estimateur sans biais de m. Calculer son risque quadratique.
3°) Comparer les deux estimateurs.

1°) $E(T_n) = m$, et $r(T_n) = V(T_n) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{m^2}{n}$.

2°) Soit F la fonction de répartition de X, et donc de toutes les variables X_i :
$(S)\begin{cases} F(x) = 0 & \text{si } x < 0 \\ F(x) = 1 - e^{-\frac{x}{m}} & \text{si } x \geq 0 \end{cases}$
$\forall x \in \mathbb{R}^+$

$$\begin{aligned} F_Y(x) &= P(Y \leq x) = 1 - P(Y > x) \\ &= 1 - P((X_1 > x) \cap (X_2 > x) \cap \ldots \cap (X_n > x)) \\ &= 1 - \prod_{i=1}^{n}(1 - F(x)) \\ &= 1 - \prod_{i=1}^{n} e^{-\frac{x}{m}} \\ &= 1 - e^{-\frac{n}{m}x} \end{aligned}$$

$\forall x \in \mathbb{R}^{-*} \qquad F_Y(x) = 1 - \prod_{i=1}^{n}(1 - 0) = 0$
donc $Y \hookrightarrow \mathcal{E}\left(\frac{n}{m}\right)$, d'où $E(Y) = \frac{m}{n}$ et $E(Z) = nE(Y) = m$.
Donc Z est un estimateur sans biais de m.
Son risque quadratique est : $r(Z) = V(Z) = n^2 V(Y) = n^2\frac{m^2}{n^2} = m^2$

3°) T_n est un meilleur estimateur que Z car son risque quadratique est inférieur à celui de Z.

3 Estimateur et estimation ponctuelle

Définition 5 : Une **estimation ponctuelle** d'un paramètre θ d'une variable aléatoire X est la valeur prise par un estimateur T_n pour un échantillon observé de la population étudiée.

Notation : Si $T_n = g(X_1, X_2, \ldots, X_n)$, l'estimation ponctuelle est notée :
$\hat{T}_n = g(x_1, x_2, \ldots, x_n)$, où les réels x_i sont les valeurs mesurées des X_i, c'est-à-dire les valeurs mesurées de la variable X sur les n individus de l'échantillon observé.

T_n est aléatoire, $\hat{T}_n$ est une valeur mesurée.

Remarque : l'estimation ponctuelle $\hat{T}_n$ dépend de l'échantillon choisi : on obtient une valeur différente avec un autre échantillon.

Par exemple : On a construit en Turbo-Pascal une simulation d'une variable X de loi hypergéométrique. La moyenne de 10000 exemplaires de cette variable nous donne un estimateur de l'espérance de X.
En lançant l'exécution du programme, on obtient une certaine valeur de cet estimateur, qui est une *estimation* de $E(X)$. Si on recommence, on va obtenir une autre estimation, etc.

En fait, on n'a aucune certitude quant au fait que l'estimation donne la *vraie* valeur de θ, ni même qu'elle en constitue une bonne valeur approchée.
On complète donc cette démarche par l'estimation par intervalle de confiance : on cherche un intervalle aléatoire qui contienne θ avec une probabilité minimale donnée.

4 Intervalle de confiance

4.1 Définition, généralités

Définition 6 :

Soit un n-échantillon $(X_1, X_2, \ldots, X_n)$ d'une variable X,
$U_n = g_1(X_1, X_2, \ldots, X_n)$, $\quad V_n = g_2(X_1, X_2, \ldots, X_n)$.
Alors $[U_n, V_n]$ est un intervalle de confiance de θ au niveau de confiance $1-\alpha$ si et seulement si $P(U_n \leq \theta \leq V_n) \geq 1-\alpha$.

En pratique : On obtient, pour un échantillon observé, des valeurs $x_1, x_2, \ldots, x_n$ mesurées. On en déduit des valeurs observées de U_n et V_n : $\hat{U}_n = g_1(x_1, x_2, \ldots, x_n)$ et $\hat{V}_n = g_2(x_1, x_2, \ldots, x_n)$.
L'intervalle $[\hat{U}_n, \hat{V}_n]$ est une estimation de l'intervalle de confiance au risque α. (C'est-à-dire au niveau de confiance $1-\alpha$).

A présent, comment déterminer les variables aléatoires U_n et V_n ?
On cherche le plus souvent cet intervalle sous la forme $[T_n - \epsilon, \ T_n + \epsilon]$, où T_n est un estimateur de θ :
$P(T_n - \epsilon \leq \theta \leq T_n + \epsilon) = P(|T_n - \theta| \leq \epsilon) = P(\theta - \epsilon \leq T_n \leq \theta + \epsilon)$

On utilisera essentiellement le **théorème de la limite centrée**, ou **l'inégalité de Bienaymé-Tchebychev**.

4.2 Avec Bienaymé-Tchebychev

Si T_n est un estimateur sans biais de θ, alors $E(T_n) = \theta$ et donc :

$$P(|T_n - \theta| \leq \epsilon) > 1 - \frac{V(T_n)}{\epsilon^2}$$

donc si on connaît $V(T_n)$, on peut calculer ϵ de sorte que $1 - \dfrac{V(T_n)}{\epsilon^2} \geq 1-\alpha$.

Cas particulier : si le paramètre à estimer est celui d'une loi de Bernoulli, noté p, on prend pour estimateur sans biais T_n la variable aléatoire : $T_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, et dans ce cas $V(T_n) = \dfrac{p(1-p)}{n}$.
On a alors :
$1 - \dfrac{V(T_n)}{\epsilon^2} \geq 1-\alpha \quad \Leftrightarrow \quad \dfrac{p(1-p)}{n\epsilon^2} \leq \alpha \quad \Leftrightarrow \quad \epsilon^2 \geq \dfrac{p(1-p)}{n\alpha}$
Or on sait que $p(1-p) \leq \dfrac{1}{4}$ d'où : $\dfrac{p(1-p)}{n\alpha} \leq \dfrac{1}{4n\alpha}$
Il suffit donc de prendre $\epsilon = \dfrac{1}{2\sqrt{n\alpha}}$:

$\left[T_n - \frac{1}{2\sqrt{n\alpha}},\ \ T_n + \frac{1}{2\sqrt{n\alpha}}\right]$ est un intervalle de confiance de p au niveau de confiance $1-\alpha$ (ou : au risque α).

4.3 Avec le théorème de la limite centrée

Exemple 5 : On suppose que $X \hookrightarrow \mathcal{N}(m;4)$.
Trouver un intervalle de confiance de m au niveau de confiance 0,95, et en déduire une estimation de cet intervalle de confiance, sachant que la moyenne des valeurs de X observée sur un échantillon d'effectif 100 est de 12,7.

Prenons un 100-échantillon $(X_1, X_2, \ldots, X_{100})$, et soit $Y_{100} = \frac{1}{100}\sum_{k=1}^{100} X_k$.

On sait que Y_{100} est un estimateur de m.
D'autre part, avec l'hypothèse d'indépendance des éléments de l'échantillon, on sait d'après le théorème de la limite centrée que Y_{100} suit approximativement une loi normale de paramètres $E(Y_{100})$ et $V(Y_{100})$.
On a : $E(Y_{100}) = m$, et $V(Y_{100}) = \frac{1}{10000}(100 \times 4) = \frac{1}{25}$.
Cherchons un réel ϵ tel que $P(Y_{100} - \epsilon \leq m \leq Y_{100} + \epsilon) = 0,95$. (On aura ainsi un intervalle de confiance de m au niveau de confiance 0,95).

$$\begin{aligned}
P(Y_{100} - \epsilon \leq m \leq Y_{100} + \epsilon) = 0,95 &\Leftrightarrow P(|Y_{100} - m| \leq \epsilon) = 0,95 \\
&\Leftrightarrow P(m - \epsilon \leq Y_{100} \leq m + \epsilon) = 0,95 \\
&\Leftrightarrow \Phi\left(\frac{m+\epsilon-m}{\frac{1}{5}}\right) - \Phi\left(\frac{m-\epsilon-m}{\frac{1}{5}}\right) = 0,95 \\
&\Leftrightarrow 2\Phi(5\epsilon) - 1 = 0,95 \\
&\Leftrightarrow \epsilon = \frac{1,96}{5} = 0,392
\end{aligned}$$

donc on a : $U_{100} = Y_{100} - 0,392$ et $V_n = Y_{100} + 0,392$.
Estimation de cet intervalle de confiance : en remplaçant Y_{100} par son estimation 12,7 on obtient :

$$[12,308; 13,092]$$

Généralisation :
Soit à estimer par intervalle de confiance au risque α l'espérance m d'une variable aléatoire X, de variance σ^2.
On pose $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$: T_n est un estimateur sans biais de $m = E(X)$.

Pour n assez grand, la loi de T_n peut être approchée par une loi normale de paramètres m et $\frac{\sigma^2}{n}$. On a alors :

$$P(T_n - \epsilon \leq m \leq T_n + \epsilon) = P(m - \epsilon \leq T_n \leq m + \epsilon) \simeq 2\Phi\left(\frac{\epsilon}{\frac{\sigma}{\sqrt{n}}}\right) - 1$$

Cherchons alors ϵ tel que $P(T_n - \epsilon \leq m \leq T_n + \epsilon) = 1 - \alpha$:

$$2\Phi\left(\frac{\epsilon}{\frac{\sigma}{\sqrt{n}}}\right) - 1 = 1 - \alpha \quad \Leftrightarrow \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right) = 1 - \frac{\alpha}{2}$$

Posons $t_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. Alors l'équation précédente s'écrit : $\frac{\epsilon\sqrt{n}}{\sigma} = t_\alpha \quad \Leftrightarrow \quad \epsilon = t_\alpha \frac{\sigma}{\sqrt{n}}$

Théorème 3 :

$\left[T_n - t_\alpha \frac{\sigma}{\sqrt{n}},\ \ T_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de m au niveau de confiance $1-\alpha$ (ou : au risque α), si $t_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

Cas particulier :
Replaçons dans le cas du paragraphe précédent, où on choisit pour estimateur d'un paramètre p d'une variable de Bernoulli X la moyenne de n exemplaires de X.
Dans la formule précédente, on a : $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$.
On peut donc prendre comme intervalle de confiance au risque α :

$$\left[T_n - \frac{t_\alpha}{2\sqrt{n}},\ \ T_n + \frac{t_\alpha}{2\sqrt{n}}\right]$$

Exemple 6 : On lance n fois une pièce telle que la probabilité d'obtenir "pile" est p, inconnu. On suppose $n > 30$.
A l'aide du théorème de la limite centrée, déterminer un intervalle de confiance de p au risque 0,01.
On effectue 100 lancers, et on obtient 54 fois pile. La pièce est-elle truquée ou équilibrée ?

Soit $T_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ l'estimateur de p, alors approximativement, $T_n \hookrightarrow \mathcal{N}(p, \frac{p(1-p)}{n})$, donc :

$$\begin{aligned} P(|T_n - p| \leq \epsilon) = 0,99 \quad &\Leftrightarrow \quad P\left(-\frac{\epsilon}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{T_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\epsilon}{\sqrt{\frac{p(1-p)}{n}}}\right) = 0,99 \\ &\Leftrightarrow \quad 2\Phi\left(\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) = 1,99 \\ &\Leftrightarrow \quad \frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}} = \Phi^{-1}(0,995) = 2,57 \end{aligned}$$

donc un intervalle de confiance au risque 0,01 de p est :

$$\left[T_n - 2,57\sqrt{\frac{p(1-p)}{n}}; T_n + 2,57\sqrt{\frac{p(1-p)}{n}}\right]$$

De plus, bien qu'on ne connaisse pas la valeur de p, on sait que $p(1-p) \leq \frac{1}{4}$, donc $\quad 2,57\sqrt{\frac{p(1-p)}{n}} \leq \frac{2,57}{2\sqrt{n}}$.
Par conséquent la probabilité que p appartienne à l'intervalle $\left[T_n - \frac{2,57}{2\sqrt{n}}; T_n + \frac{2,57}{2\sqrt{n}}\right]$ sera supérieure à 0,99 : on peut dire que ce dernier intervalle est un intervalle de confiance au risque 0,01 de p.
En prenant $n = 100$ et la valeur donnée de l'estimation de T_n, on obtient une estimation de cet intervalle de confiance :

$$[0,4115; 0,6685]$$

Conclusion : La probabilité que p soit dans cet intervalle est supérieure à 0,99.
Comme la valeur 0,5 appartient aussi à cet intervalle, on ne peut pas conclure que la pièce est truquée, on ne sait pas non plus si elle est équilibrée !
Pour pouvoir aller plus loin il faudrait un plus grand nombre de lancers, par exemple 10000, ou alors augmenter le risque de se tromper (prendre par exemple un niveau de confiance de 0,9 au lieu de 0,99).