Estimation

Problème : Soit une variable aléatoire réelle Y dont la loi de probabilité $\mathcal{L}_\theta$ dépend d'un paramètre réel inconnu noté θ. On cherche à utiliser l'information contenue dans un échantillon pour répondre à l'une des deux questions :

1) Quelle est la valeur du paramètre θ ? C'est le problème de l'estimation ponctuelle.
2) Dans quelle "fourchette" peut-on situer θ ? C'est le problème de l'estimation par intervalle.

Estimation ponctuelle

Définition 1 *1 - Un échantillon de taille N de Y est un ensemble de N variables aléatoires $(Y_1, ..., Y_N)$ indépendantes et de même loi que Y (on écarte le cas où les Y_n ne sont pas indépendantes).*

2 - Un ensemble $(y_1, ..., y_N)$ de réalisations des N variables Y_n est une observation, ou réalisation de l'échantillon. C'est un point de $\mathbb{R}^N$.

3 - Si $\mathcal{L}_\theta$ est une loi discrète, sa distribution étant $L(y;\theta) = P_\theta\{Y = y\}$, la loi de distribution de l'échantillon est

$$\begin{aligned} L(y_1, ...y_N;\theta) &= P\{Y_1 = y_1, ..., Y_N = y_N;\theta\} \\ L(y_1, ...y_N;\theta) &= P_\theta\{Y_1 = y_1\} P_\theta\{Y_2 = y_2\} ...P_\theta\{Y_N = y_N\} \end{aligned}$$

4 - Si $\mathcal{L}_\theta$ est une loi continue de densité $f(y;\theta)$, la loi de distribution de l'échantillon est :

$$L(y_1, ..., y_N;\theta) = f(y_1;\theta)f(y_2;\theta)...f(y_N;\theta)$$

5 - Dans les deux cas, la fonction $L(y_1, ..., y_N;\theta)$ considérée comme fonction de θ s'appelle vraisemblance du paramètre (attachée à l'observation de $y_1, ..., y_N$).

6 - Un estimateur de θ est une variable aléatoire $T_N = t(Y_1, ...Y_N)$, où t est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ qui, aux valeurs observées $(y_1, ...y_N)$, fait correspondre le nombre réel :

$$\widehat{\theta}_N = t(y_1, ..., y_N)$$

dit estimation de θ.

Remarque 2 *l'estimateur $T_N = t(Y_1, ...Y_N)$, fonction de N variables aléatoires, est une variable aléatoire (si f est une fonction dite "mesurable", ce qui sera le cas pour toutes les fonctions envisagées ici).*

l'estimation $\widehat{\theta}_N = t(y_1, ..., y_N)$ est un nombre certain, réalisation de la variable aléatoire T_N.

Ces définitions sont encore valables si le paramètre est constitué de k coordonnées. La fonction t est alors une fonction de $\mathbb{R}^N$ dans $\mathbb{R}^k$.

A - Estimateurs sans biais, estimateurs convergents

Définition 3 *Un estimateur T_N est sans biais si son espérance mathématique est la vraie valeur du paramètre :*

pour tout θ, $E_\theta(T_N) = \theta$

Définition 4 *Un estimateur T_N est convergent en probabilité s'il tend en probabilité vers θ lorsque N augmente infiniment.*

Rappel : "T_N tend en probabilité vers θ" s'écrit $T_N \xrightarrow{P} \theta$,
et signifie que pour tout $\varepsilon > 0$, $P\{|T_N - \theta| < \varepsilon\} \longrightarrow 1$ lorsque N tend vers l'infini.

Théorème 5 *CS de convergence en probabilité : $\{E(T_N) \longrightarrow \theta \text{ et } V(T_N) \longrightarrow 0\}$.*

Un estimateur dont l'espérance mathématique tend vers θ et dont la variance tend vers 0 est convergent.

Remarque 6 *a fortiori, un estimateur sans biais dont la variance tend vers zéro est convergent.*

B - Comparaison de deux estimateurs sans biais

Un estimateur sans biais est plus précis, donc meilleur, qu'un autre estimateur sans biais si sa variance est plus petite.

Définition 7 *La quantité d'information contenue dans l'échantillon, relative au paramètre θ est*

$$I_N(\theta) = E\left[\left(\frac{\partial \ln L\left(Y_1,..,Y_N;\theta\right)}{\partial\theta}\right)^2\right]$$

Théorème 8 *Inégalité de CRAMER-RAO. Si l'intervalle de variation de la variable aléatoire Y ne dépend pas du paramètre à estimer, et sous réserve de certaines conditions de régularité de la fonction de vraisemblance*[1]*, la variance d'un estimateur sans biais T_N est bornée inférieurement. Elle satisfait à l'inégalité :*

$$V(T_N) \geq \frac{1}{I_N(\theta)}$$

L'hypothèse précédente écarte en particulier les lois uniformes sur l'intervalle $[0;\theta]$. Dans ce cas, l'estimateurdu maximum de vraisemblance a une variance plus petite que la borne de Cramer-Rao (exemple traité en TD de DEUG Sciences Economiques).

Théorème 9 *Sous les hypothèses de Cramer-Rao, la quantité d'information peut aussi se calculer par une formule plus facile à calculer :*

$$\begin{aligned} I_N(\theta) &= V\left[\frac{\partial \ln L\left(Y_1,..,Y_N;\theta\right)}{\partial\theta}\right] \\ I_N(\theta) &= -E\left[\frac{\partial^2 \ln L\left(Y_1,..,Y_N;\theta\right)}{\partial\theta^2}\right] \end{aligned}$$

Définition 10 *Un estimateur sans biais est efficace si sa variance est égale à* $\frac{1}{I_N(\theta)}$

C - Méthode du maximum de vraisemblance : Estimation ponctuelle d'un paramètre

On choisit comme estimation $\widehat{\theta}_N$ celle qui pour $(y_1,...,y_N)$ fixés rend maximum la vraisemblance du paramètre. En général[2], l'estimation du maximum de vraisemblance $\widehat{\theta}_N$ vérifiera :

$$\begin{cases} CN1 : \frac{\partial L(y_1,..,y_N;\theta)}{\partial\theta} = 0 \\ CS2 : \frac{\partial^2 L(Y_1,..,Y_N;\theta)}{\partial\theta^2} < 0 \end{cases}$$

ou de façon équivalente :

$$\begin{cases} CN1 : \frac{\partial \ln L(y_1,..,y_N;\theta)}{\partial\theta} = 0 \\ CS2 : \frac{\partial^2 \ln L(Y_1,..,Y_N;\theta)}{\partial\theta^2} < 0 \end{cases}$$

Théorème 11 *Sous réserve des conditions d'application de l'inégalité de CRAMER-RAO (certaines conditions de régularité et si l'intervalle de variation de Y ne dépend pas de θ), l'estimateur T_N du maximum de vraisemblance :*

a) tend en probabilité vers la vraie valeur du paramètre .

b) est asymptotiquement Normal, sans biais et efficace.

[1] Toutes les lois rencontrées dans le cours de Licence satisfont à ces conditions de régularité.

[2] la première égalité n'est nécessaire que si la vraisemblance est continue et dérivable en θ et si le maximum n'est pas sur le bord de l'ensemble de définition. La seconde condition n'est que suffisante.

Cela signifie que pour N assez grand, on peut approcher la loi de T_N par une loi Normale d'espérance θ et de variance $\frac{1}{I_N(\theta)}$.

Théorème 12 *Si il existe un estimateur efficace, il est solution de l'équation du maximum de vraisemblance.*

Remarque 13 *Ces deux théorèmes justifient l'emploi fréquent de cette méthode.*

Les propriétés décrites dans le premier théorème étant des propriétés asymptotiques (ie quand $N \longrightarrow \infty$) la méthode sera d'autant mieux justifiée que la taille de l'échantillon sera grande.

D : Estimation simultanée de plusieurs paramètres par la méthode du maximum de vraisemblance.

La variable aléatoire Y suit une loi de probabilité $\mathcal{L}_\theta$ dépendant d'un paramètre qui peut être un ensemble de plusieurs nombres réels : $\theta = (\theta_1, .., \theta_p)$

La vraisemblance du paramètre est alors $L(y_1, ...y_N; \theta_1, .., \theta_p)$.

La méthode du maximum de vraisemblance conduit en général à la résolution du système d'équations :

$$CN1 \quad : \quad \begin{cases} \frac{\partial L(y_1,..,y_N;\theta)}{\partial \theta_1} = 0 \\ \vdots \\ \frac{\partial L(y_1,..,y_N;\theta)}{\partial \theta_p} = 0 \end{cases}$$

$$CS2 \quad : \quad \left[\frac{\partial^2 L\,(Y_1, .., Y_N; \theta)}{\partial \theta_i \partial \theta_j}\right]_{i,j=1,..p} \quad \text{matrice définie négative}$$

Attention : les conditions suffisantes pour que la solution trouvée corresponde à un maximum font intervenir le tableau complet des dérivées secondes de la log-vraisemblance [3]

E - Méthode des moments

La variable aléatoire Y suit une loi de probabilité $\mathcal{L}_\theta$ dépendant d'un paramètre (qui peut être un ensemble de plusieurs nombres réels, $\theta = (\theta_1, .., \theta_p)$. Nous ne supposons pas que cette loi est entièrement déterminée par la donnée de θ. Nous supposons seulement qu'il est possible de calculer en fonction de θ les p premiers moments de Y, notés $m_1, m_2, \ldots, m_p$. Nous supposons également que ces p équations permettent de calculer θ en fonction des moments de Y :

$$\begin{cases} E(Y) =: m_1 = f_1\,(\theta_1, .., \theta_p) \\ E(Y^2) =: m_2 = f_2\,(\theta_1, .., \theta_p) \\ \vdots \\ E(Y^p) =: m_p = f_p\,(\theta_1, .., \theta_p) \end{cases} \Longleftrightarrow \begin{cases} \theta_1 = g_1\,(m_1, .., m_p) \\ \theta_2 = g_2\,(m_1, .., m_p) \\ \vdots \\ \theta_p = g_p\,(m_1, .., m_p) \end{cases}$$

L'estimateur de θ par la méthode des moments est obtenu en remplaçant dans ces expressions les moments théoriques par les moments empiriques de Y calculés à partir d'un échantillon de taille N :

$$\begin{cases} \widehat{m_1} = \frac{1}{N} \sum\limits_{n=1}^{N} Y_n \\ \widehat{m_2} = \frac{1}{N} \sum\limits_{n=1}^{N} Y_n^2 \\ \vdots \\ \widehat{m_p} = \frac{1}{N} \sum\limits_{n=1}^{N} Y_n^p \end{cases} \longmapsto \begin{cases} \widehat{\theta_1} = g_1\,(\widehat{m_1}, .., \widehat{m_p}) \\ \widehat{\theta_2} = g_2\,(\widehat{m_1}, .., \widehat{m_p}) \\ \vdots \\ \widehat{\theta_p} = g_p\,(\widehat{m_1}, .., \widehat{m_p}) \end{cases}$$

Si les fonctions g_i sont continues en $(m_1, m_2, \ldots, m_p)$, alors les estimateurs obtenus sont des estimateurs convergents.

[3] cf. par exemple "Fonctions de plusieurs variables" de BOUZITAT - PRADEL, Cujas.

Exemple 14 *Un assureur étudie la distribution du coût d'un certain sinistre pour un certain type d'assuré. Un modèle classique attribue à ce coût Y une loi Gamma, dont la densité dépend de deux paramètres strictement positifs, a et r :*

$$L(y; a, r) = \frac{y^{r-1}}{a^r \Gamma(r)} e^{-y/a} \quad \text{pour } y > 0, \quad 0 \text{ sinon}$$

On peut montrer que $E(Y) = ra$, et $V(Y) = ra^2$. On en tire immédiatement que

$$\begin{cases} a = \frac{V(Y)}{E(Y)} \\ r = \frac{[E(Y)]^2}{V(Y)} \end{cases}$$

et les estimateurs par la méthode des moments sont :

$$\begin{cases} \widehat{a} = \frac{S^2}{\overline{Y}} \\ \widehat{r} = \frac{\overline{Y}^2}{S^2} \end{cases} \quad \text{où } \overline{Y} = \frac{1}{N}\sum_{n=1}^{N} Y_n \quad \text{et } S^2 = \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \overline{Y}\right)^2$$

F - Méthode d'ajustement par les moindres carrés ordinaires(MCO)

Nous cherchons ici à expliquer la variable Y en fonction de variables (dites explicatives) : $x1, x2, .., xp$. Nous avons N observations $(Y_1, x1_1, x2_1, .., xp_1), .., (Y_N, x1_N, x2_N, .., xp_N)$.

Pour simplifier les notations, nous notons $x_n =: (x1_n, x2_n, .., xp_n)$ l'ensemble des variables explicatives.

Définition 15 *Dans la famille des fonctions $f(x;\theta)$, où $\theta \in \Theta$, la fonction ajustée par les Moindres Carrés Ordinaires (MCO) est la fonction "la plus proche" des valeurs observées, au sens où :*

$$\widehat{f} = f(x; \widehat{\theta}), \ \widehat{\theta} \text{ solution de } \underset{\theta \in \Theta}{Min} \sum_{n=1}^{N} [y_n - f(x_n;\theta)]^2$$

Cas de l'ajustement à une droite passant par l'origine y = ax

Calcul de l'estimateur

Les N observations (y_n, x_n) étant fixées, la fonction objectif est $Q(a) = \sum_{n=1}^{N} [y_n - ax_n]^2$

Les dérivées d'ordre 1 et 2 de $Q(a)$ sont :

$$\begin{cases} \frac{\partial Q(a)}{\partial a} = 2\sum_{n=1}^{N}(y_n - ax_n)(-x_n) = -2\sum_{n=1}^{N}(y_n - ax_n)x_n \\ \frac{\partial^2 Q(a)}{\partial a^2} = 2\sum_{n=1}^{N} x_n^2 > 0 \end{cases}$$

($\frac{\partial^2 Q(a)}{\partial a^2}$ est strictement positive si les x_n ne sont pas tous nuls).

La condition nécessaire d'ordre 1 entraîne donc

$$\widehat{a} = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}$$

qui correspond bien à un minimum de la fonction $Q(a)$, puisque la condition suffisante d'ordre 2 est satisfaite.

Propriétés statistiques dans le cadre du modèle linéaire standard : si nous supposons que

$$\begin{aligned} H1 &: E(y_n) = ax_n \\ H2 &: V(y_n) = \sigma^2, Cov(y_n, y_m) = 0 \text{ si } n \neq m \end{aligned}$$

l'estimateur $\widehat{a}$ a pour espérance et pour variance :

$$\begin{aligned} E(\widehat{a}) &= \frac{\sum_{n=1}^{N} E(y_n) x_n}{\sum_{n=1}^{N} x_n^2} = \frac{\sum_{n=1}^{N} a x_n^2}{\sum_{n=1}^{N} x_n^2} = a \\ V(\widehat{a}) &= \frac{\sum_{n=1}^{N} V(y_n) x_n^2}{\left[\sum_{n=1}^{N} x_n^2\right]^2} = \frac{\sigma^2}{\sum_{n=1}^{N} x_n^2} \end{aligned}$$

La condition de convergence de $\widehat{a}$ vers a est donc que la somme des carrés des x_n tende vers l'infini.

Cas de l'ajustement à une droite y = ax + b

Calcul de l'estimateur :

Les N observations (y_n, x_n) étant fixées, la fonction objectif est $Q(a,b) = \sum_{n=1}^{N} [y_n - ax_n - b]^2$

Les dérivées d'ordre 1 et 2 de $Q(a,b)$ sont :

$$\left\{ \begin{array}{l} \frac{\partial Q(a,b)}{\partial a} = 2 \sum_{n=1}^{N} (y_n - ax_n - b)(-x_n) = -2 \sum_{n=1}^{N} (y_n - ax_n - b) x_n \\ \frac{\partial Q(a,b)}{\partial b} = 2 \sum_{n=1}^{N} (y_n - ax_n - b)(-1) = -2 \sum_{n=1}^{N} (y_n - ax_n - b) \\ \frac{\partial^2 Q(a,b)}{\partial a^2} = 2 \sum_{n=1}^{N} x_n^2 \quad \frac{\partial^2 Q(a,b)}{\partial a \partial b} = 2 \sum_{n=1}^{N} x_n \\ \qquad\qquad\qquad\qquad \frac{\partial^2 Q(a,b)}{\partial b^2} = 2 \sum_{n=1}^{N} 1 = 2N \end{array} \right.$$

Les conditions nécessaires d'ordre 1 (appelées « équations normales ») sont :

$$\left\{ \begin{array}{l} \sum_{n=1}^{N} (y_n - ax_n - b) x_n = 0 \\ \sum_{n=1}^{N} (y_n - ax_n - b) = 0 \end{array} \right.$$

Nous avons un système de deux équations pour les deux inconnues (a,b) :

$$\left\{ \begin{array}{l} a \sum_{n=1}^{N} x_n^2 + b \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} y_n x_n \\ a \sum_{n=1}^{N} x_n + bN = \sum_{n=1}^{N} y_n \end{array} \right.$$

Nous notons

$$\begin{aligned}
\text{moments empiriques} &: \quad \overline{x} = \tfrac{1}{N}\sum_{n=1}^{N} x_n \quad \overline{y} = \tfrac{1}{N}\sum_{n=1}^{N} y_n \\
Var_N(x) &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2 \\
Cov_N(y,x) &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})(y_n - \overline{y})
\end{aligned}$$

Rappelons les formules classiques de statistique descriptive

$$\begin{aligned}
\sum_{n=1}^{N}(x_n - \overline{x}) &= 0 \text{ et } \sum_{n=1}^{N}(y_n - \overline{y}) = 0 \\
Var_N(x) &= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - (\overline{x})^2 \\
Cov_N(y,x) &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})\, y_n = \frac{1}{N}\sum_{n=1}^{N} x_n y_n - \overline{xy}
\end{aligned}$$

La solution des équations normales est, si les x_n ne sont pas constantes :

$$\begin{cases} \widehat{a} = \frac{Cov_N(y,x)}{Var_N(x)} \\ \widehat{b} = \overline{y} - \widehat{a}\overline{x} \end{cases}$$

La matrice des dérivées secondes est $Hess = 2\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix}$.

C'est une matrice 2x2, sa trace est > 0 et son déterminant est

$$\det(Hess) = 4\left[N\sum_{n=1}^{N} x_n^2 - \left(\sum_{n=1}^{N} x_n\right)^2\right] = 4N^2 Var_N(x) > 0$$

Les deux valeurs propres de H sont de même signe positif. La condition suffisante d'ordre 2 est satisfaite : la solution trouvée correspond bien à un minimum de la fonction $Q(a,b)$

Théorème 16 *Propriétés statistiques dans le cadre du modèle linéaire standard :*
$si \begin{cases} H1 : E(y_n) = ax_n + b \\ H2 : V(y_n) = \sigma^2 \text{ et } Cov(y_n, y_m) = 0 \text{ si } n \neq m \end{cases}$ *, alors :*

$$\begin{aligned}
E[\widehat{a}] &= a, \ E\left[\widehat{b}\right] = b \\
V(\widehat{a}) &= \sigma^2 \frac{1}{\sum_{n=1}^{N}(x_n - \overline{x})^2}, \ V\left(\widehat{b}\right) = \sigma^2 \frac{\frac{1}{N}\sum_{n=1}^{N} x_n^2}{\sum_{n=1}^{N}(x_n - \overline{x})^2} = \sigma^2\left(\frac{1}{N} + \frac{(\overline{x})^2}{\sum_{n=1}^{N}(x_n - \overline{x})^2}\right) \\
Cov\left(\widehat{a}, \widehat{b}\right) &= -\sigma^2 \frac{\overline{x}}{\sum_{n=1}^{N}(x_n - \overline{x})^2}
\end{aligned}$$

Nous constatons que l'estimateur $\widehat{a}$ est convergent si et seulement si $\sum_{n=1}^{N}(x_n - \overline{x})^2 \longrightarrow \infty$.