MATHÉMATIQUES ET SCIENCES HUMAINES

M. TABET

L'estimation

Mathématiques et sciences humaines, tome 5 (1964), p. 23-26

<http://www.numdam.org/item?id=MSH_1964__5__23_0>

M. TABET

L'ESTIMATION

I - ESTIMATION PONCTUELLE

Supposons que l'on tire un échantillon $(x_1, x_2, \ldots, x_n)$ d'une population dont la distribution a une forme mathématique connue, mais fait intervenir un paramètre inconnu θ. L'échantillon doit nous permettre de calculer un nombre que l'on pourra prendre pour la valeur de θ, ce sera l'estimation $t\ (x_1, \ldots, x_n)$ de θ. Or les valeurs $x_1, \ldots, x_n$ sont les valeurs prises par une variable aléatoire X. Il est bien entendu que l'estimation t est une fonction des valeurs $x_1, x_2, \ldots, x_n$; comme celles-ci sont aléatoires de loi de probabilité connue, il convient donc de considérer l'estimation t comme la valeur prise par une certaine variable aléatoire T.

La variable aléatoire T est appelée un estimateur, ses valeurs t sont des estimations.

Plus précisément, l'estimateur constitue la règle qui associe au paramètre θ à estimer une fonction d'éléments observables donc aléatoires, issus de la population et en nombre fini: $T\ (X_1, X_2, \ldots, X_n)$. Les réalisations $(x_1, x_2, \ldots, x_n)$ donnent la valeur $t\ (x_1, x_2, \ldots, x_n)$ qui est l'estimation ponctuelle de θ, ponctuelle parce qu'elle associe une unique valeur connue t, à une valeur inconnue θ.

On notera que souvent estimation prend le sens d'estimateur lorsqu'il n'y a pas de risque d'interprétation erronée.

Par exemple, on estimera la moyenne (inconnue) θ d'une v.a. X de fonction de répartition F (x) (de forme connue: par exemple, Laplace-Gauss d'écart type unité) par l'estimateur: "moyenne arithmètique $\frac{x_1 + x_2 + \ldots + x_n}{n}$ des n valeurs $x_1, x_2, \ldots, x_n$ trouvées dans l'échantillon". L'estimateur est cette règle et, $x_1, x_2, \ldots, x_n$ étant des réalisations particulières de v.a. $X_1, X_2, \ldots, X_n$ (toutes de loi F(x)), c'est la v.a. $\frac{X_1 + X_2 + \ldots + X_n}{n}$, dont la loi de probabilité se déduit de celle, F, de X. (De façon analogue la régle qui a u associe la valeur $f(u) = u^2$ est la fonction f, les valeurs qu'elle prend sont les nombres u^2).

Le problème est non pas de trouver des estimations, mais des estimateurs. Cependant, le fait qu'un estimateur soit une variable aléatoire signifie que l'on ne peut prédire les estimations dans un cas particulier, mais seulement à la longue, en moyenne. Il est donc évident qu'un estimateur qui donne dans un seul cas particulier une valeur t très différente de θ ne saurait être rejeté. Pour que l'estimateur T convienne il suffit que sa distribution ne soit pas trop dispersée autour de θ.

D'après ce qui précéde on voit qu'un estimateur n'est défini qu'assez vaguement. Pour choisir parmi tous les estimateurs possibles on leur imposera des conditions.

II - QUALITES D'UN ESTIMATEUR

1) Un estimateur T sans distorsion ou sans biais est tel que $E(T) = \theta$.

Par exemple, si θ est l'espérance mathématique ou moyenne théorique d'une variable aléatoire X, l'estimateur $T = \frac{\sum X}{n}$ vérifie $E(T) = \frac{\sum E(X)}{n} = E(X) = \theta$.

2) Un estimateur T convergent est tel que la probabilité pour que les valeurs de T soient éloignées de θ devient arbitrairement petite lorsque la taille n de l'échantillon augmente. Autrement dit, les valeurs de T sont très proches de θ, sauf dans une proportion négligeable de cas:

$$Pr\ (|T - \theta| < \varepsilon) > 1 - \eta$$

ε et η étant rendus aussi petits que l'on désire en prenant n suffisamment grand.

Pratiquement ceci revient à dire que lorsque T a une variance, celle-ci doit tendre vers 0 lorsque la taille n de l'échantillon tend vers l'infini (T est d'autant plus concentré autour de θ que l'échantillon est plus grand).

$V(T) < \eta$ aussi petit que soit η, dès que n est assez grand.

Par exemple pour $T = \frac{\sum X}{n}$

$$V(T) = V\left(\frac{\sum X}{n}\right) = \frac{1}{n} V(X) \quad \text{(si X a une variance)}$$

Or $\frac{1}{n} V(X) < \eta$ dès que n est assez grand, aussi petit que soit η. Si par exemple $V(X) = 2 = 10^{-4}$, on aura $\frac{1}{n} V(X) < \eta$ dès que $n \geqslant 2.10^4$.

Un estimateur est absolument correct s'il est à la fois sans biais et convergent.

3) Un estimateur absolument correct T sera d'autant plus efficace que sa variance V(T) sera plus petite: cela signifie en effet que plus la variance est petite plus l'estimateur T est concentré autour de son espérance θ. Il faut noter que s'il existe un biais, la notion d'efficacité doit être précisée.

On montre que, dans certaines conditions, il existe un estimateur absolument correct T^* de variance minimum: il est dit efficace. T^* est, en un certain sens, le "meilleur" des estimateurs possibles. Un tel estimateur, de variance minimum existe en particulier pour la loi de Laplace-Gauss, la loi de Poisson, la loi binomiale.

4) Un estimateur T est exhaustif s'il résume toute l'information contenue dans l'échantillon $x_1, x_2, \ldots, x_n$ et relative au paramètre à estimer.

III - FONCTION DE VRAISEMBLANCE

Soit $x_1, x_2, \ldots, x_n$ un échantillon tiré d'une population dont la distribution a une densité f. La fonction de vraisemblance est par définition

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

soit $\log. L(x, \theta) = \sum_i \log. f(x_i, \theta)$

On estimera alors θ par la valeur $\hat{\theta}$ qui rend $L(x, \theta)$, ou son logarithme, maximum.

$$L(x, \hat{\theta}) = \max L(x, \theta)$$

Si l'intervalle de variation de θ ne dépend pas de la valeur θ, l'estimation $\hat{\theta}$ vérifie, lorsque f est dérivable par rapport à θ :

$$\frac{\delta \log L}{\delta \theta}(x, \hat{\theta}) = 0$$

Donc on prendra $\hat{\theta}$ de façon à vérifier l'équation précédente.

Si par contre, l'intervalle de variation de θ dépend de la valeur θ à estimer, on ne peut plus appliquer cette méthode.

Par exemple: soit à estimer sur un échantillon de n valeurs la longueur a de l'intervalle $(0, a)$ d'une distribution uniforme. On a $f(x)\,dx = \frac{dx}{a}$

$$L(x, a) = \frac{1}{a^n} \quad \text{et} \quad \max L(x, a) = L(x, 0)$$

On devrait donc estimer a par 0, ce qui est absurde. Dans ce cas particulier d'ailleurs, on peut estimer a en utilisant la plus grande valeur de l'échantillon: $\hat{a}$ (max x_j).

Remarque: parfois on cherche un estimateur linéaire.

$T = \sum_i \alpha_i X_i$ où $\sum_i \alpha_i = 1$, les α_i étant inconnus.

On détermine alors les α_i par les conditions

$$\begin{cases} E(T) = \theta \\ V(T) \text{ minimum} \end{cases}$$

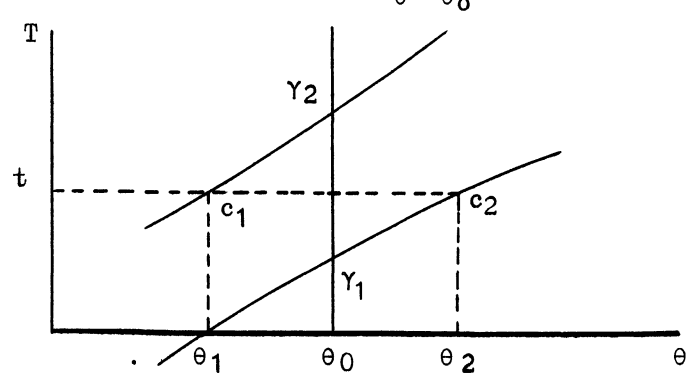
IV - ESTIMATION PAR INTERVALLE

Au lieu d'estimer la valeur d'un paramètre θ par un nombre unique t, on se propose de trouver un intervalle I, tel que I contienne θ avec une probabilité $1 - \alpha$ donnée d'avance: P (I contient θ) $= 1 - \alpha$

Soit donc T un estimateur de θ, prenant la valeur t pour l'échantillon $x_1, x_2, \ldots x_n$.

Pour une valeur θ_0 de θ on peut déterminer deux nombres γ_1 et γ_2 dépendants de θ_0 et α tels que:

$$P(\gamma_1 < t < \gamma_2)/_{\theta = \theta_0} = 1 - \alpha \quad \text{(cf. fig.)}$$



I

Lorsqu'on considère toutes les valeurs possibles de θ, les points γ_1 et γ_2 correspondant à chacune de ces valeurs décrivent deux courbes (cf. fig.).

Ainsi on a défini pour chaque valeur de θ un intervalle (γ_1, γ_2) dans lequel la variable aléatoire T a une probabilité $1 - \alpha$ de se trouver et par complémentarité une probabilité α d'être en dehors.

La réalisation t de l'estimateur T étant apparue, on peut effectuer une partition sur l'ensembledes valeurs possibles de θ en distinguant les valeurs de θ :

1) pour lesquelles t appartient à l'intervalle $(\gamma_1, \gamma_2)_\theta$

2) pour lesquelles t n'appartient pas à l'intervalle $(\gamma_1, \gamma_2)_\theta$

On voit que graphiquement cela revient à tracer l'horizontale passant par t qui coupe les courbes de γ_1 et γ_2 en c_2 et c_1 d'où les valeurs θ_2 et θ_1 constituant les bornes de l'intervalle I.

L'intervalle (θ_1, θ_2) constitue l'intervalle de confiance du paramètre θ. La probabilité que cet intervalle recouvre la vraie valeur θ, certaine mais non connue, est $1 - \alpha$. Les bornes θ_1 et θ_2 sont déterminées par la réalisation t d'une variable aléatoire T : elles sont donc aléatoires en ce sens qu'elles procèdent de T.

BIBLIOGRAPHIE

KENDALL: The advanced theory of statistics
Londres - Griffin

CRAMER: Mathematical Methods of Statistics
Princeton University Press.