Institut National de la Statistique et des Etudes Economiques

74

ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ÉCONOMIQUE

STATISTIQUE NON PARAMETRIQUE

par

J.L. PIEDNOIR

DIRECTION GÉNÉRALE

3, Avenue Pierre Larousse - 92240 MALAKOFF Téléphone : 540-12-12

COURS

de

STATISTIQUE NON PARAMETRIQUE

(J.L. PIEDNOIR)

Maître-assistant à l'Université de Paris-Sorbonne

Professeur à l' E.N.S.A.E.

9038.79.1R

PLAN DU COURS

COURS

de

STATISTIQUE NON PARAMETRIQUE

(J.L. PIEDNOIR)

Maître-assistant à l'Université de Paris-Sorbonne

Professeur à l' E.N.S.A.E.

CHAPITRE 0

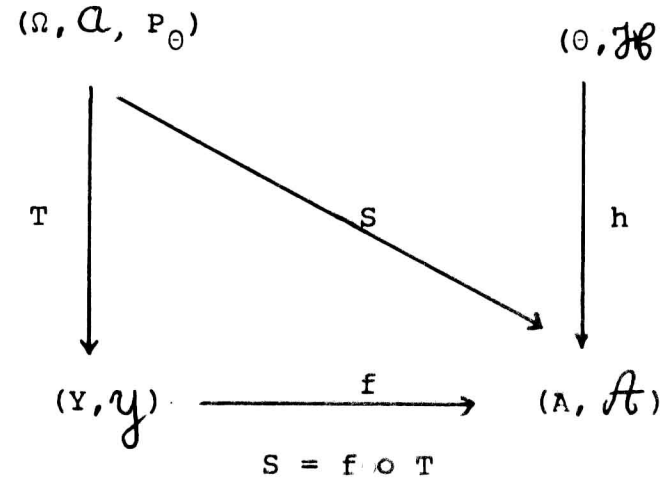
EN GUISE D'INTRODUCTION

0.1. QU'EST-CE QUE LA STATISTIQUE NON PARAMETRIQUE ?

Tout problème de statistique inférentielle (non séquentiel) peut se formaliser de la façon suivante : on a un constat expérimental ω et $\omega \in \Omega$. Ω est appelé ensemble des cas possibles, c'est l'ensemble des constats à priori possibles pour l'expérience analysée ; Ω est un espace probabilisé noté $(\Omega, \mathcal{A}, P)$ et on suppose, c'est l'hypothèse fondamentale, que la loi de probabilité P caractérise le phénomène naturel qui a abouti au constat ω . Mais la loi P est inconnue, on sait seulement à priori que $P \in (P_\theta)_{\theta \in \Theta}$, famille de lois de probabilité, $(P_\theta)_{\theta \in \Theta} \subset \mathcal{P}$ ensemble des lois de probabilité sur Ω . Pour cette raison P sera noté P_θ et il existe $\theta \in \Theta$ caractérisant le phénomène étudié. On désire avoir des renseignements sur ce θ particulier au phénomène, ce qui se formalise de la façon suivante : on se donne une application $h : \Theta \to A$, on veut "connaître" $h(\theta)$.

L'objet de la statistique inférentielle est en un certain sens de "mesurer" $h(\theta)$ à partir du constat ω donc d'être capable de choisir $a \in A$ à partir de ω . Une stratégie du statisticien sera donc une application $S : \Omega \to A$, stratégie en général non déterminée directement mais à partir d'une statistique T application de Ω dans Y . Tous les espaces étant supposés mesurables et les applications décrites également, on a le schéma suivant :

.../...



Dans la plupart des problèmes classiques, $\Omega = X^N$ avec $X = \mathbb{R}^k$.

$P_\theta = \prod_{i=1}^{N} \pi_{i,\theta}$ où $\pi_{i,\theta}$ est une loi de probabilité sur X et si μ est la mesure de Lebesgue sur $\mathbb{R}^k$ on suppose que $\frac{d\pi_{i,\theta}}{d\mu} = f_i(x,\theta)$ et que Θ est un borélien d'intérieur non vide de $\mathbb{R}^p$. On dit alors que le problème dépend de p paramètres réels (l'application $\theta \rightsquigarrow f_i(x,\theta)$ est injective).

Exemple : Le problème des deux échantillons.

C'est le problème qui nous servira à illustrer de nombreux cas dans la suite de ce cours. On suppose que : $X = \mathbb{R}$

$$\frac{d\pi_{i,\theta}}{d\mu}(x) = f(x,\theta_1,\sigma) \quad i \leq m$$

$$\frac{d\pi_{i,\theta}}{d\mu}(x) = f(x,\theta_2,\sigma) \quad m+1 \leq i \leq m+n = N \qquad \theta = (\theta_1,\theta_2,\sigma)$$

.../...

On a donc $\Theta = \mathbb{R}^2 \times \mathbb{R}^+$. On veut tester l'hypothèse $\theta_1 = \theta_2$ contre l'alternative $\theta_1 \neq \theta_2$. On a donc $A = \{0,1\}$ et

$$h(\theta) = 0 \quad \text{si} \quad \theta_1 = \theta_2$$

$$h(\theta) = 1 \quad \text{si} \quad \theta_1 \neq \theta_2$$

f est supposé connu, l'inférence porte sur $\theta_1 - \theta_2$.

Un tel modèle dit paramétrique étant posé, les techniques statistiques classiques recherchent les statistiques T sur lesquelles on peut fonder des stratégies présentant certains caractères d'optimalité.

Mais la validité d'un tel modèle n'est pas toujours assurée. En général il est difficile de choisir un type de densité; supposer que celle-ci est gaussienne par exemple est souvent plus une hypothèse avancée pour des raisons de simplicité des techniques à mettre en place qu'une hypothèse fondée en raison. Que se passe-t-il si ce modèle n'est plus valide ?

Pour répondre à la question on a le choix entre deux directions de recherche :

1) On munit $\mathcal{P}$ ensemble des lois de probabilité sur Ω d'une topologie adéquate et on suppose que la loi P du phénomène est dans un voisinage de la famille $(P_\theta)_{\theta \in \Theta}$ et on cherche des procédures qui, si elles ne sont plus optimales pour la famille $(P_\theta)_{\theta \in \Theta}$, gardent de bonnes propriétés d'optimalité dans un voisinage précisé de cette famille. C'est le problème de la robustesse. Dans l'exemple du problème des deux échantillons, on définit une distance sur les lois de probabilité sur $\mathbb{R}$ (distance de Prokhorow) on prend sur $\mathcal{P}$ la topologie produit et on prend un voisinage produit de la famille $(f(x,\mu,\sigma))_{\mu,\sigma \in \mathbb{R} \times \mathbb{R}^+}$

.../...

2) On abandonne l'idée d'un paramétrage possible par $\mathbb{R}^p$ et on se place d'emblée dans un ensemble plus grand de lois de probabilité, c'est la voie non paramétrique. Dans l'exemple du problème des deux échantillons on suppose f inconnue et on cherche des procédures qui soient encore valides. Dans ce cas particulier, on peut encore reformuler le problème de la façon suivante. On suppose :

$$\frac{d\pi_{i,\theta}}{d\mu}(x) = f(x) \quad i \leq m$$

$$\frac{d\pi_{i,\theta}}{d\mu}(x) = f(x-\Delta)$$

On teste $\Delta = 0$ contre $\Delta \neq 0$

L'espace des paramètres est alors $\Theta = \mathbb{R} \times \mathcal{F}$ où $\mathcal{F}$ est l'ensemble des densités de probabilité sur $\mathbb{R}$ supposées continues. Si la densité f est continue, f est définie par les valeurs $(f(x))_{x \in \mathbb{Q}}$. Comme $\mathbb{Q}$ ensemble des rationnels est dénombrable on voit que :

$\mathcal{F} = \mathbb{R}^{\mathbb{N}}$ noté en général $\mathbb{R}^{\infty}$.

Plus généralement on pourrait définir un problème non paramétrique comme un problème où l'ensemble des paramètres est du type $\mathbb{R}^{\infty}$. Malheureusement cette définition est inadéquate pour des problèmes séquentiels justiciables des techniques classiques. Par exemple on suppose la suite de populations gaussiennes d'écart-type 1 et de moyenne μ_i avec $i \in \mathbb{N}$. Ici le paramétrage naturel est encore $\mathbb{R}^{\infty}$.

On se contente donc de la définition floue suivante : on appellera problème non paramétrique un problème dans lequel l'ensemble des lois de probabilité à priori possible est un espace très général.

.../...

Ce qui bien entendu ne veut pas dire qu'il n'y a pas de paramètres, mais que l'espace des paramètres n'est plus un espace de dimension fini comme $\mathbb{R}^p$.

0.2. NOTATIONS

Au long de ce cours nous emploierons les notations suivantes :

- Ω sera l'espace fondamental, ensemble de tous les résultats à priori possible.

- Si Ω est un espace produit on notera : $\Omega = X^N$

- $\mathcal{X}$ sera la tribu des événements sur X, $\mathcal{A}$ la tribu des événements sur Ω avec $\mathcal{A} = \mathcal{X}^{\otimes N}$

- P sera une loi de probabilité sur Ω, Π une loi de probabilité de X

- Soit μ une mesure σ - finie sur Ω ; si P admet une densité par rapport à μ, celle-ci sera notée p .

- $\mathcal{P}$ sera l'ensemble des lois de probabilité sur $(\Omega, \mathcal{A})$

- $\mathcal{P}_\mu$ sera l'ensemble des lois de probabilité sur $(\Omega, \mathcal{A})$ admettant une densité par rapport à μ .

- $\mathcal{P}(H)$ sera l'ensemble des lois de probabilité sur $(\Omega, \mathcal{A})$ satisfaisant une hypothèse H .

- Si T est une statistique à valeurs dans $(Y, \mathcal{Y})$; T application mesurable de Ω dans T, la sous-tribu engendrée par T sera nommée $\mathcal{A}^T = T^{-1}(\mathcal{Y})$.

- Q étant une loi de probabilité sur $(Y, \mathcal{Y})$, On dit que T suit la loi Q noté $T \rightsquigarrow Q$ si et seulement si (noté si ssi) $\forall B \in \mathcal{Y}$ $Q(B) = P[T^{-1}(B)]$.

- Une suite de problèmes sera notée $(\Omega_\nu, \mathcal{A}_\nu, P_\nu)_{\nu \in N}$

.../...

La suite de statistiques $(T_\nu)_{\nu\in\mathbb{N}}$, $T_\nu:\Omega_\nu\to(Y,\mathcal{Y})$, suit asymptotiquement la loi Q si ssi :

$$\forall B\in\mathcal{Y}\quad P_\nu[T_\nu^{-1}(B)]\underset{\nu\to\infty}{\to}Q(B)$$

Ceci sera noté $T_\nu \underset{\nu\to\infty}{\rightsquigarrow} Q$

On définit également les fonctions numériques suivantes :

$$u(.)\ :\ u(x)=1\quad x\geqslant 0$$
$$u(x)=0\quad x<0$$

$$s(.)\ :\ s(x)=1\quad x>0$$
$$s(o)=0$$
$$s(x)=-1\quad x<0$$

- Considérons l'ensemble $\{1,2,\ldots,n\}$ qui sera noté E_n . L'ensemble des bijections de E_n , ensemble des permutations des entiers $(1,2,\ldots,n)$ sera noté $\mathfrak{S}_n$.

0.3. DEFINITIONS

- Statistique libre

- Soit $(\Omega,\mathcal{A},P)$ un espace fondamental, T une statistique $T:(\Omega,\mathcal{A})\to(Y,\mathcal{Y})$; T est dite libre par rapport à la famille $\mathcal{D}$ de lois de probabilité sur $(\Omega,\mathcal{A})$ si et seulement si :

$\exists Q$ loi de probabilité sur $(Y,\mathcal{Y})$ tel que :

$$\forall P\in\mathcal{D}\quad Q=T^{-1}(P)$$

Autrement dit la loi de probabilité de T est indépendante de $P\in\mathcal{D}$.

.../...

- Statistique complète

- Une famille de lois de probabilité $\mathcal{D}$ est dite complète (ou totale) si et seulement si f étant une variable aléatoire telle que :

$$\forall P \in \mathcal{D} \quad E_P(f) = 0 \text{ , alors } \forall P \in \mathcal{D} \quad f = 0 \text{ P.p.p.}$$

- Une statistique T est dite complète sur la famille $\mathcal{D}$ de lois de probabilité si et seulement si la famille de lois de probabilité $\mathcal{D}^T$ est complète, $\mathcal{D}^T$ étant l'ensemble des lois P^T, où P^T est P restreint à la tribu $\mathcal{A}^T$.

- On notera que si T est exhaustive et complète sur $\mathcal{D}$ alors $\mathcal{D}$ est complète.

- Statistique bornée complète

- Même définition que précédemment avec la restriction f bornée.

- Problème invariant

- Soit $\mathcal{G}$ un groupe de transformations sur $(\Omega, \mathcal{A})$ tel que la famille $(P_\theta)_{\theta \in \Theta}$ soit fermée pour $\mathcal{G}$: $P_\theta \circ g^{-1} \in (P_\theta)_{\theta \in \Theta}$

- On définit $\bar{g} : \Theta \to \Theta$ par :

$$P_{\bar{g}(\theta)} = P_\theta \circ g^{-1} \text{ ; } \bar{g} \text{ est bijectif.}$$

L'ensemble $\bar{\mathcal{G}}$ des transformations $\bar{g}$ est un groupe.

- Le problème statistique est dit invariant par $\mathcal{G}$ si et seulement si $\forall a \in A \quad \forall \theta_1 \in h^{-1}(\{a\}), \forall \theta_2 \in h^{-1}(\{a\}) \; \forall \bar{g} \in \bar{\mathcal{G}}$

$$h[\bar{g}(\theta_1)] = h[\bar{g}(\theta_2)]$$

.../...

- On peut alors définir une transformation $\bar{\bar{g}}$ sur A par :

$$\bar{\bar{g}}(a) = h[\bar{g}(\theta)] \quad \theta \in h^{-1}(\{a\})$$

- Une stratégie S est dite invariante si :

$$\forall g \in \mathcal{G} \quad Sog = \bar{\bar{g}}oS$$

- <u>Statistique d'ordre, statistique de rang</u>

- Soit $X_1 \ldots X_N$ N variables aléatoires. On pose :

$$R_i = \sum_{j=1}^{N} u(X_i - X_j)$$

R_i est le rang de X_i parmi $X_1 \ldots X_N$

$R = (R_1 \ldots R_N)$ est le vecteur des rangs , $R(\omega) \in \mathfrak{S}_N$

- On pose :

$$X^{(i)} = X_j \text{ si et seulement si } i = R_j$$

On a $X^{(1)} \leq X^{(2)} \leq \ldots \leq X^{(N)}$

Les $X^{(1)} \ldots X^{(N)}$ sont les $X_1 \ldots X_n$ ordonnés par valeurs croissantes ; $X^{(.)} = (X^{(1)}, \ldots, X^{(N)})$ est la statistique d'ordre.

Il est équivalent de se donner $X = (X_1, \ldots, X_N)$ ou $(X^{(.)}, R)$

- On posera également :

$$R_i^+ = \sum_{j=1}^{N} u(|X_i| - |X_j|)$$ rang de $|X_1|$ parmi $|X_1| \ldots |X_N|$

$|X|^{(.)}$ sera la statistique d'ordre de $|X_1| \ldots |X_N|$

$$S_i = s(X_i)$$

$$S = (S_1, \ldots, S_N)$$

Il est équivalent de se donner X ou $(|X|^{(.)}, R, S)$

.../...

0.4. RAPPEL DE LA THEORIE DES TESTS

0.4.1. Problème de test

- Un problème statistique est un problème de test si $A = \{0,1\}$
- Si on admet les décisions randomisées on appellera fonction test $\phi : \Omega \to [0,1]$; $\phi(\omega)$ est la probabilité de décider 1 si le résultat expérimental est ω .
- Si on suit la formalisation de Neymann - Pearson le test ϕ est de niveau α si $E_\theta(\phi) = \int_\Omega \phi(\omega)\, P_\theta(d\omega) \leq \alpha \quad \forall \theta \in h^{-1}(\{0\}) = \theta_0$
- On considère $M(\alpha, \mathcal{P}(H_0))$ ensemble des tests de niveau α pour l'hypothèse $H_0 : \theta \in \Theta_0$

La fonction puissance du test ϕ sera la quantité :

$$\beta_\phi(\theta) = E_\theta(\phi) \qquad \theta \in h^{-1}(\{1\}) = \Theta_1$$

Un test est sans biais si :

$$\beta_\phi(\theta) \geq \alpha \qquad \forall \theta \in h^{-1}(\{1\})$$

- Un test ϕ de niveau α est dit UMP sur $\mathcal{P}(H_1)$ si et seulement si $\forall \psi \in M(\alpha, P_0)$ $\forall P \in \mathcal{P}_1$ $\beta_\phi(P) \geq \beta_\psi(P)$

- Un test ϕ de niveau α est dit UMPU sur $\mathcal{P}(H_1)$ s'il est UMP parmi les test sans biais

- Si Θ est un espace topologique et $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \Theta \setminus \{\theta_0\}$, un test ϕ de niveau α sera dit localement UMP (ou localement UMPU) si et seulement si il existe un voisinage V de θ_0 tel que ϕ est UMP (ou UMPU) dans ce voisinage.

- Un test ϕ est dit à niveau constant si et seulement si:

$$\forall \theta \in \Theta_0 \quad E_\theta(\phi) = \alpha$$

.../...

0.4.2. Test à structure de Neymann

Définition : Un test ϕ est dit à structure de Neymann par rapport à une statistique T si et seulement si :

$$\forall P \in \mathcal{P}(H_o) \quad E_p(\phi|T) = \alpha \quad P\text{-p.p}$$

Théorème de Lehmann : Soit T une statistique exhaustive sur $\mathcal{P}(H_o)$.

Les deux propriétés suivantes sont équivalentes :

(I) T est bornée complète sur $\mathcal{P}(H_o)$

(II) Tout test ϕ à niveau constant est à structure de Neymann par rapport à T.

Démontrons (I) => (II) : Soit ϕ un test à niveau constant

$$\forall P \in \mathcal{P}(H_o) \quad E_p(\phi) = \alpha \quad \text{donc}$$

$$E_p[E_p(\phi-\alpha|T)] = E_p(\phi-\alpha) = 0$$

T étant exhaustive $E_p(\phi-\alpha|T)$ est indépendant de P et on pose $\phi(T) = E(\phi-\alpha|T)$

On a $E_p[\phi(T)] = 0 \qquad \forall P \in \mathcal{P}(H_o)$

T étant borné complète sur $\mathcal{P}(H_o)$ ceci implique que :

$$\phi(T) = 0 \quad \mathcal{P}(H_o)\text{ - p.p.} \qquad \text{et donc}$$

$$E_p(\phi|T) = \alpha \quad \mathcal{P}(H_o)\text{ - p.p.}$$

Démontrons que (II) => (I) : Pour cela on montrera que non (I) => non (II)

Si T n'est pas bornée complète sur $\mathcal{P}(H_o)$, T étant exhaustive, il existe une fonction f sur $(Y, \mathcal{Y})$ bornée telle que $\forall P \in \mathcal{P}(H_o) E_p[f(T)] = 0$ et $\exists P \in \mathcal{P}(H_o)$ tel que $f(.) \neq 0$ P-p.p.

Posons $M = \sup_{t \in Y} f(t)$ $\quad C = \min(\frac{\alpha}{M}, \frac{1-\alpha}{M})$ $\quad \phi(t) = C\, f(t) + \alpha, \phi(t) \in [0,1]$

$\phi(T)$ est une fonction test telle que

$$\forall P \in \mathcal{P}(H_o) \quad E_p[\phi(T)] = \alpha$$

donc ϕ est à niveau constant mais ϕ n'est pas à structure de Neymann.

.../...

TITRE 1

LES TESTS NON PARAMETRIQUES