STATISTIQUE EXHAUSTIVE (G1, G5)
(26 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

Notion fondamentale de la **Statistique** (R.A. FISHER), une **statistique « exhaustive »** possède la propriété de contenir toute l'**information** (utile) pour réaliser une **inférence statistique** relative à un **paramètre d'intérêt**.

(i) Soit $((\mathcal{X}, \mathcal{B}), (P_\theta^X)_{\theta \in \Theta})$ un **modèle image** exprimé sous forme paramétrée, $(\mathcal{S}, \mathcal{E})$ un **espace mesurable** auxiliaire, $s : \mathcal{X} \mapsto \mathcal{S}$ une **application mesurable** (ou **statistique**), et S = s (X) la « statistique » déduite de s. On peut présenter le concept de statistique exhaustive de trois façons.

En effet, on dit que s (resp S) est une **statistique exhaustive** pour le paramètre $\theta \in \Theta$, ou pour la **famille** de **lp** $(P_\theta^X)_{\theta \in \Theta}$, ssi :

(a) ou bien $s^{-1} (\mathcal{E}) \subset \mathcal{B}$ est une sous-**tribu exhaustive** de $\mathcal{B}$. On note parfois σ (s) ou σ (S) cette sous-tribu.

Autrement dit, S (ω) = s o X (ω) étant connue (observée dans $\mathcal{S}$), la connaissance de ω lui-même n'informe pas davantage sur $\theta \in \Theta$, ie sur la « vraie » **loi de probabilité** P_θ^X de X ;

(b) ou bien la **loi conditionnelle** de X sachant S ne dépend pas de θ. On dit encore que s (resp S) est exhaustive pour $\theta \in \Theta$ ssi, quelle que soit la statistique z (resp Z) à valeurs dans un espace $(\mathcal{Z}, \mathcal{D})$ et définie par Z = z o X, la loi conditionnelle de z (resp Z) sachant S est indépendante de θ. On retrouve ce qui précède en posant $z = id_{\mathcal{X}}$ et en supposant que $(\mathcal{Z}, \mathcal{D}) = (\mathcal{X}, \mathcal{B})$;

(c) ou bien (P.R. HALMOS-L.J. SAVAGE) il existe une détermination de la loi conditionnelle de X sachant s (resp S) qui est indépendante de $\theta \in \Theta$. Autrement dit, il existe dans ce cas une fonction mesurable $s \mapsto Q (B / S = s)$ indépendante de $\theta \in \Theta$ et tq (cf **application mesurable**) :

$$(1) \qquad P_\theta^X (B \cap s^{-1} (E)) \qquad \text{ou} \qquad P_\theta (X \in B] \cap [S \in E]) = \int_E Q (B / S = s)\, dP_\theta^S (s),$$

pour tout couple de **parties** $(B, E) \in \mathcal{B} \times \mathcal{E}$ et tout $\theta \in \Theta$. On note $(P_\theta^S)_{\theta \in \Theta}$ la **famille des lois** qui sont les images des lois P_θ^X par s (ie $P_\theta^S = s (P_\theta^X)$), ou les images des **mesures de probabilité** P_θ par S = s o X (ie $P_\theta^S = s \circ X (P_\theta) = P_\theta^{s \circ X}$) (cf **mesure image**), et l'on note par s aussi bien la statistique $s : \mathcal{X} \mapsto \mathcal{E}$ que l'une de ses valeurs s o X (ω) = S $(\omega) \in \mathcal{E}$.

Il existe ainsi une **probabilité de transition** N de $\mathcal{X} \times \mathcal{E}$ dans $\mathcal{E}$ tq :

(1)' $P_\theta^X(B \cap s^{-1}(E))$ ou $P_\theta([X \in B] \cap [S \in E]) = \int_C N(s, B)\, dP_\theta^S(s)$,

pour tout couple $(B, E) \in \mathcal{B} \times \mathcal{E}$ et tout $\theta \in \Theta$.

(ii) Une statistique exhaustive possède d'importantes propriétés :

(a) elle « apporte » sur θ la même « **quantité d'information** » (au sens de FISHER) que l'**échantillon** dont elle provient (cf **information de FISHER**). En ce sens, elle ne génère pas de **perte d'information** ;

(b) elle existe toujours (eg l'échantillon X lui-même) ;

(c) si $s : \mathcal{X} \mapsto \mathcal{E}$ et $t : \mathcal{X} \mapsto \mathcal{Z}$ sont deux statistiques et si $\phi : \mathcal{E} \mapsto \mathcal{Z}$ est une **application bimesurable** (ie bijective et mesurable, ainsi que son inverse), alors :

(2) {s est exhaustive} $\Rightarrow$ {$t = \phi \circ s$ est exhaustive} ;

(d) si $\{(\mathcal{X}, \mathcal{B}), (P_\theta^X)_{\theta \in \Theta}, (D, \mathcal{B}_D), L\}$ est un **problème de décision** statistique associé au modèle précédent, et si s (resp S) est une statistique exhaustive pour θ, alors l'ensemble $\mathcal{C}_s$ des **règles de décision pures** $\delta_s : \mathcal{X} \mapsto D$ basées sur s forme une classe essentiellement complète (ie $\mathcal{C}_s$ est formé des éléments extrémaux pour la **relation d'ordre** définie sur D par la **fonction de risque** R associée à la **fonction de perte** L du problème) (cf **classe complète de règles de decision**).

(iii) Une statistique exhaustive peut se définir aussi bien sur un modèle image (comme ci-dessus) que sur le modèle de base $((\Omega, \mathcal{T}, (P_\theta)_{\theta \in \Theta})$ qui le sous-tend. Dans le second cas, le formalisme des notations est plus simple.

(iv) Si X est un N-échantillon $(X_1, ..., X_N)$ de taille N potentiellement variable, la propriété d'exhaustivité conserve son intérêt ssi elle est vérifiée quelle que soit N. Si c'est le cas, on dit parfois que X « permet » une statistique exhaustive de θ. Il suffit, pour celà, que cette propriété soit vérifiée à partir d'une certaine taille $N \in \mathbf{N}^*$.

(v) On peut s'intéresser à un **paramètre dérivé** $\tau = g(\theta)$ au lieu du paramètre $\theta \in \Theta$ lui-même, avec $g : \Theta \mapsto \Lambda$ (**application mesurable** donnée, à valeurs dans un espace $(\Lambda, \mathcal{B}_\Lambda)$). La notion d'exhaustivité s'étend directement à cette situation. Il en est de même en présence de **paramètre importun**.

(vi) Une statistique exhaustive permet souvent de « simplifier » un **problème statistique** lorsque la **variable** observée X est une variable « complexe » à analyser (eg lorsque le nombre N des observations X_n est élevé). Le **critère de factorisation** permet, sous certaines hypothèses, d'exhiber de telles statistiques.

Pour traiter un problème statistique, le **principe d'exhaustivité** consiste ainsi à n'utiliser que des statistiques exhaustives.

A l'opposé d'une statistique exhaustive, une **statistique libre** (ou **statistique ancillaire**) n'apporte aucune information sur le paramètre d'intérêt $\theta \in \Theta$. **Exhaustivité** et **liberté** sont donc deux concepts « antinomiques ».

(vii) Une terminologie ancienne qualifiait une statistique exhaustive s (resp S) de **« résumé exhaustif »** (G. DARMOIS), voire de « *statistique suffisante* » (de l'anglais « *sufficient* statistic »), pour désigner une statistique exhaustive, ie une statistique qui épuise toute l'information apportée par l'échantillon.