THÉORIQUE, EMPIRIQUE (C, F)
(28 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La distinction entre **théorique** et **empirique** est une distinction de base permettant de séparer les notions (cf notamment **statistique naturelle**) :

(a) de **concept théorique**, qui est relative à un **ensemble** fondamental Ω (eg **population** ou « population mère »). A cette notion se rattache la notion de **loi de probabilité** « théorique » ;

(b) de **concept empirique**, qui est relative à un sous-ensemble A (eg **échantillon**) de la population Ω précédente. A ce niveau se rattache la notion de **loi de probabilité** « empirique » (cf **loi empirique**).

On peut aussi distinguer (notamment en **théorie des sondages**) entre :

(a) la **loi** selon laquelle on « tire » les **unités statistiques** $a \in A$ dans la population Ω, qui est la loi Π des **unités de sondage** (cf **plan de sondage**) ;

(b) la loi résultant de ce tirage, et solidarisant les **variables** observables sur ces unités, qui est la loi P^{η} des **variables aléatoires** η associées à ces dernières.

Ces notions s'entendent donc toujours au sens du **calcul des probabilités**.

La distinction précédente ne correspond pas nécessairement à celle faite entre **Statistique inférentielle** (ou **Statistique « mathématique »**) et **Statistique descriptive** : en effet, cette dernière peut porter aussi bien sur l'échantillon A que sur la population Ω elle-même.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $(\mathcal{Z}, \mathcal{D})$ un **espace d'observation** et $\zeta : \Omega \mapsto \mathcal{Z}$ une **variable aléatoire** (a priori) **observable** qui se relie à (ie décrit) un **phénomène** donné.

Soit, par ailleurs, $Z = (Z_1, ..., Z_N) : \Omega \mapsto \mathcal{Z}^N$ un **vecteur aléatoire** (ou une **suite** finie) à valeurs dans l'espace produit $\mathcal{Z}^N$, et constituée de N va quelconques Z_n (cf **échantillon aléatoire**). Les N coordonnées Z_n de Z (échantillon de valeurs) peuvent être :

(a) non indépendantes entre elles (cas général) (cf **dépendance**), auquel cas la loi de Z est définie comme l'image $P^Z = Z(P)$ de P par Z ;

(b) indépendantes entre elles (situation « intermédiaire ») (cf **indépendance stochastique**), auquel cas la loi de Z est de la forme $P^Z = \otimes_{n=1}^{N} P^{Z(n)}$ (produit tensoriel), où Z(n) désigne Z_n (n = 1 ,..., N) et $P^{Z(n)}$ désigne la loi marginale (ou loi « propre ») de Z_n ($\forall$ n $\in N_N^*$) (cf **échantillon indépendant**) ;

(c) équidistribuées comme une **variable parente** ζ, ie $Z_n \sim P^\zeta$, $\forall\ n \in N_N^*$ (situation « intermédiaire »), auquel cas la loi P^Z de Z est tq ses marginales se confondent avec une loi fixe de la forme $P^{Z(0)}$, ie $P^{Z(n)} = P^{Z(0)}$, $\forall\ n \in N_N^*$ (cf **échantillon équidistribué**) ;

(d) indépendamment équidistribuées comme ζ, auquel cas $P^Z = \otimes_{n=1}^N P^\zeta = (P^\zeta)^{\otimes N}$ (puissance tensorielle de P^ζ) (cf **échantillon iid**).

(ii) On doit donc distinguer trois types de **loi de probabilité** :

(a) la loi $P^\zeta = \zeta(P)$ de ζ (image de P par ζ), appelée **loi de probabilité théorique** ou **loi théorique** (situation d'équidistribution). On suppose souvent que $P^\zeta \in \mathscr{P}^\zeta$ (ensemble de **lp** définies sur $\mathscr{D}$) ;

(b) la **loi (de probabilité) de l'échantillon** Z (ou d'une **statistique** dépendant de Z, cf infra) : par définition, cette loi est l'image $P^Z = Z(P)$ de P par Z. Elle est généralement appelée **loi d'échantillon(nage)** ou **distribution d'échantillon(nage)**. On suppose souvent que $P^Z \in \mathscr{P}^Z$ (ensemble de lois possibles pour Z) ;

(c) la **loi empirique**, associée à l'échantillon Z selon :

(1) $\quad P_N = N^{-1} . \Sigma_{n=1}^N \delta(Z_n)$,

où $\delta(z)$ désigne la **loi de DIRAC** placée au point $z \in \mathscr{Z}$, avec $Z = (Z_1, ..., Z_N)'$. P_N est une loi de probabilité elle-même « aléatoire », ou « loi stochastique », puisqu'elle dépend de Z. En particulier, si Z est un **échantillon iid** comme la variable parente ζ, on a $P^Z = (P^\zeta)^{\otimes N}$. Cette loi est toujours calculée comme en (1), quelle que soit la nature de ζ ou de Z : elle ne parcourt donc pas, a priori, un ensemble potentiel de lois empiriques.

De même, si $(\mathscr{S}, \mathscr{E})$ désigne un **espace probabilisable** auxiliaire et $s : \mathscr{Z} \mapsto \mathscr{Y}$ une **application mesurable** définissant une **statistique** $S = s(Z) : \Omega \mapsto \mathscr{S}$, alors S possède une loi de probabilité définie par $\mathscr{L}(S) = S(P) = (s \circ Z)(P) = s(P^Z)$. Cette loi est parfois aussi appelée **loi d'échantillon(nage)** ou **distribution d'échantillon(nage)** : c'est la loi d'une statistique définie à partir de Z.

(iii) Soit $\gamma = c(P^\zeta)$ la valeur d'une **caractéristique** associée à chaque loi P^ζ. On note Γ l'ensemble de ces valeurs et $c : P^\zeta \mapsto \Gamma$ le mode opératoire (ou **application caractéristique**) qui associe ces caractéristiques aux lois. La **loi de probabilité** (théorique) P^ζ de ζ possède ses propres caractéristiques. On appelle alors :

(a) **caractéristique théorique** de P^ζ (ou de ζ) l'image γ de P^ζ par c, ie :

(2) $\quad \gamma = c(P^\zeta), \quad \forall\ P^\zeta \in \mathscr{P}^\zeta$.

Il s'agit d'une grandeur non aléatoire (sauf dans un contexte bayésien) (cf infra).

Ainsi :

$(a)_1$ lorsque c associe à P^ζ le **moment algébrique** non centré d'ordre j ($\gamma = \mu_j$), on définit le **moment non centré théorique** d'ordre j selon :

(3) $\mu_j = \int z^j \, dP^\zeta (z)$;

$(a)_2$ lorsque c associe à P^ζ le p-**quantile** ($\gamma = Q_p \, \zeta$), on définit le **quantile théorique** d'ordre $p \in]0, 1[$ (supposé unique), selon :

(4) $H (Q_p \, \zeta) = p$,

où H est la **fonction de répartition** associée à P^ζ (avec ici $\mathcal{Z} =$ **R**) ;

(b) **caractéristique empirique**, aussi appelée **caractéristique d'échantillon**, définie de façon analogue et calculable à partir de Z selon :

(5) γ_N ou $g_N = c (P_N)$ (image de la loi empirique par c).

Il s'agit d'une grandeur aléatoire, car elle dépend des N valeurs de $Z = (Z_1 ,..., Z_N)$ dont la loi est $P^Z = Z (P)$. Autrement dit, ζ est observée sur N **unités statistiques** a_n (n = 1 ,..., N) « participant » au phénomène considéré, avec $Z_n = \zeta (a_n)$, $\forall \, n \in N_N^*$. Lorsque Z est un **échantillon iid** issu de P^ζ, les Z_n sont des **copies** de ζ, ie $P^Z = (P^\zeta)^{\otimes N}$. La **loi empirique** P_N de Z permet alors de définir la notion de **caractéristique empirique** précédente.

Ainsi :

$(b)_1$ lorsque c associe à P_N le moment (algébrique) non centré d'ordre j, on définit le **moment non centré empirique** d'ordre j selon :

(6) $m_j = \int z^j \, dP_N (z) = N^{-1} \Sigma_{n=1}^N Z_n^j$;

$(b)_2$ de même, lorsque c associe à P_N le p-quantile, on obtient le **quantile empirique** d'ordre p, q_p Z, tq :

(7) $H_N (q_p X) = p$,

où H_N est la **fr empirique** associée à P_N .

(iv) On doit donc aussi distinguer, parallèlement, trois types de caractéristiques :

(a) la **caractéristique « théorique »**, qui se rapporte à la loi P^ζ . C'est une grandeur non aléatoire, sauf dans un cadre bayésien (cf **école bayésienne**) ;

(b) la **caractéristique « empirique »**, qui se rapporte à la loi empirique P_N ; c'est une va souvent considérée comme une statistique. C'est le cas de $g_N : \Omega \mapsto \Gamma$ définie en (5) ;

(c) la **caractéristique d'échantillonnage**, qui se rapporte à une **loi d'échantillonnage** (tq eg P^Z) ou, plus généralement, à la loi d'une statistique (tq eg $\mathscr{L}(S)$). Une caractéristique d'échantillonnage n'est pas, en général, aléatoire (même exception qu'en (a)) ;

(v) Dans ce qui précède, les coordonnées Z_n de Z ne sont pas nécessairement indépendantes, mais leurs lois marginales doivent être identiques (égales à P^ζ). On a indiqué que la situation non équidistribuée (avec indépendance ou non) est plus complexe (cf **processus stochastique**), car la loi de Z ne peut en général s'écrire sous la forme d'un **produit tensoriel** $P^Z = (P^\zeta)^{\otimes N}$.

(vi) La **distinction** précédente entre (a) caractéristique théorique, (b) caractéristique empirique et (b) caractéristique d'échantillonnage se déduit ainsi de celle existant entre les lois. C'est par le moyen de ces diverses caractéristiques que l'on peut pratiquer l'**inférence statistique** (eg estimation, tests).

La loi empirique permet le calcul de diverses **statistiques** (cf aussi **fonction des moments empiriques**). En effet, une caractéristique théorique admet généralement pour estimateur naturel la **caractéristique empirique** analogue, ie celle calculée à l'aide de P_N (cf **statistique naturelle**). Si l'opération c a un sens, une caractéristique $\gamma = c(P^\zeta)$ peut ainsi être (a priori) estimée par son analogue empirique $g_N = c(P_N)$.

Cet **estimateur** n'est pas nécessairement le meilleur et doit parfois être corrigé (cf **correction**, **modification**, **moment corrigé**, **amélioration**, etc).

(vi) Par ailleurs, certaines caractéristiques empiriques peuvent formellement ressembler à des caractéristiques théoriques (similitudes formelles, ou apparentes) : ceci est notamment le cas lorsque les premières sont calculées à partir d'une **loi uniforme discrète**. Ainsi, dans le cas général :

$$(8)\quad \begin{aligned} &E\,\zeta = \int \zeta\, dP = \int \zeta(\omega)\, dP(\omega) = \int z\, dP^\zeta(x) && \text{(\textbf{espérance} théorique)},\\ &E\,P_N = \int \zeta\, dP_N = N^{-1} \Sigma_n Z_n = e_N' Z / N = \bar{Z}_N && \text{(\textbf{moyenne empirique})}. \end{aligned}$$

Cependant, lorsque $\mathscr{Z} = N_M^* = \{1, ..., M\}$ et que ζ suit une **loi uniforme discrète** sur $N_M^* = \{1,...,M\}$, ie lorsque $P^\zeta = \mathscr{U}(N_M^*)$, l'espérance théorique vaut $E\,\zeta = e_M' z / M$, expression formellement identique à celle de la moyenne empirique précédente :

$$(9)\quad \begin{aligned} &E\,\zeta = M^{-1} . \Sigma_{m=1}^{M} z_m ,\\ &\bar{Z}_N = E\,P_N = N^{-1} . \Sigma_{n=1}^{N} z_n . \end{aligned}$$

(vi) En pratique, l'ensemble $\mathscr{Z}$ est souvent un ensemble numérique (eg **N**, **Z** ou **R**), ou une puissance de cet ensemble (eg $\mathbf{N}^K$, $\mathbf{Z}^K$ ou $\mathbf{R}^K$), ou encore, parfois, une partie de l'un des ensembles précédents.