STATISTIQUE D'ORDRE (F6)
(17 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion de **statistique d'ordre** intervient souvent en **Statistique non paramétrique**, notamment en matière de **tests d'hypothèses**, lorsqu'il est utile de classer des grandeurs (**observations**) par valeurs croissantes (ou, alternativement, décroissante).

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $X = (X_1 ,..., X_N) : \Omega \mapsto \mathbf{R}^N$ un N-**échantillon aléatoire** réel dont la **lp** est notée $P^X = X (P)$.

A chaque « réalisation » $X (\omega) = x$ de X on peut associer une **permutation** $\sigma_\omega \in \boldsymbol{\sigma}_N$ (groupe des permutations de $N_N^* = \{1 ,..., N\}$) et un **vecteur aléatoire** $(Y_1 ,..., Y_N) : \Omega \mapsto \mathbf{R}^N_{\leq}$ tq, à la fois :

(1) $\quad Y_n (\omega) \leq Y_{n+1} (\omega), \qquad \forall\, n \in \{1 ,..., N-1\},$

(2) $\quad Y_n (\omega) = X_{\sigma\omega(n)} (\omega), \qquad \forall\, n \in \{1 ,..., N\}.$

où $\sigma\omega(n)$ désigne $\sigma_\omega(n)$.

Autrement dit, Y est déterminé selon :

$$
\begin{aligned}
&Y_1 = \min \{X_n : n \in \{1 ,..., N\} \} = X_{\sigma\omega(1)} ,\\
&Y_2 = \min \{X_n : n \in \{1 ,..., N\} \setminus \{\sigma_\omega(1)\}\} = X_{\sigma\omega(2)} ,\\
&\qquad ...\\
&Y_\alpha = \min \{X_n : n \in \{1 ,..., N\} \setminus \{\sigma_\omega(1) ,..., \sigma_\omega(\alpha-1)\}\} = X_{\sigma\omega(\alpha)} ,\\
&\qquad ...\\
&Y_N = \max \{X_n : n \in \{1 ,..., N\}\} = X_{\sigma\omega(N)} .
\end{aligned}
\tag{3}
$$

On dit que Y est l'**échantillon ordonné**, ou l'**échantillon d'ordre**, associé à X et on le note $Y = X^{(.)} = (X^{(1)} ,..., X^{(N)})$, avec $X^{(n)} \leq X^{(n+1)}$, $\forall\, n \in \{1 ,..., N-1\}$ (ordre croissant).

On rencontre aussi les notations $X_{(.)}$, $X^{()}$, , $X_{()}$, $X^{:}$ ou encore $X_{:} = (X_{1:N} ,..., X_{N:N})$, avec $X_{n:N} \leq X_{n+1:N}$, $\forall\, n \in \{1 ,..., N-1\}$.

L'application $s : \mathbf{R}^N \mapsto \mathbf{R}^N_{\leq}$ définit donc une **statistique** $S = s (X)$ selon :

(4) $\quad S = s (X) = X^{(.)} = Y.$

C'est pourquoi $X^{(.)}$ est aussi appelé **statistique d'ordre**, ou **statistique ordonnée**, associée à X.

Toute statistique $T = t (X^{(.)})$ définie à partir de la statistique d'ordre $X^{(.)}$ est souvent elle-même appelée **statistique d'ordre**.

Enfin, on adopte parfois l'ordre inverse du précédent $\leq$.

(ii) Lorsque plusieurs coordonnées de X sont égales (**observation multiple**), on convient généralement d'identifier les permutations correspondantes (cf **rang multiple**).

La notion se généralise au cas où $\mathbf{R}^N$ est remplacé par un ensemble $\mathcal{X} = \mathcal{X}_0^N$ (puissance cartésienne d'un ensemble $\mathcal{X}_0$ totalement ordonné) (cf **relation d'ordre**).

D'un point de vue terminologique, on dit que :

(a) $Y_n = X_{(n)}$ est la n-ième valeur de l'échantillon $X_{(.)}$;

(b) $Y_1 = X_{(1)} = \min_n X_n$ est la (ou une) valeur minimum de X, et $Y_N = \max_n X_n$ la (ou une) valeur maximum de X. On dit encore que $X_{(1)}$ est l'**extrême** (ou l'**extrémité**) inférieur(e) de X et $X_{(N)}$ l'extrême (ou l'extrémité) supérieur(e) de X. Le couple $(X_{(1)}, X_{(N)})$ est souvent appelé **couple des (valeurs) extrêmes** de l'échantillon X ; de même le L+M-uple $(X_{(1)}, ..., X_{(L)}, ..., X_{(N-M+1)}, ..., X_{(N)})$, dans lequel $1 \leq L + M \leq N$ est appelé **statistique des valeurs extrêmes** (cf aussi **valeur extrême**, **statistique des extrêmes**) ;

(c) la différence $R_N = X_{(N)} - X_{(1)}$ est l'**étendue empirique** de X (ou des observations X_n).

(iii) La statistique d'ordre possède les propriétés suivantes :

(a) si X est un **échantillon indépendant** (non nécessairement équidistribué), alors $Y = X^{(.)}$ n'est plus indépendant, car les inégalités $Y_1 \leq ... \leq Y_N$ contraignent le support de la loi P^Y de Y : Supp $P^Y = \mathbf{R}^N_{\leq}$ (cf **support d'une probabilité**).

(b) si la loi P^X de X est une **loi absolument continue** pr à la **mesure de LEBESGUE** λ_N et si $Y = X^{(.)}$ est la statistique d'ordre de X, alors la **densité** $g = dP^Y / d\lambda_N$ de Y s'exprime en fonction de la densité $f = dP^X / d\lambda_N$ de X selon :

(5) $g(y) = \mathbf{1}_I(y) . \Sigma_{\sigma \in \mathfrak{S}(N)} f(\sigma(y))$,

où l'on note $I = \mathbf{R}^N_<$, $\sigma(y) = (y_{\sigma(1)}, ..., y_{\sigma(N)})$ et $\mathfrak{S}(N) = \mathfrak{S}_N$ est le **groupe** des permutations de N_N* déjà mentionné.

Si, de plus, X est un **échantillon iid** selon la loi P^ξ d'une **variable parente** $\xi : \Omega \mapsto \mathbf{R}$ (ie si $P^X = (P^\xi)^{\otimes N}$), la densité g de Y s'écrit :

(5)' $g(y) = N ! . 1_I(y) . \Pi_{n=1}^N f_0(y_n)$,

où $I = \mathbf{R}^N_<$, $f_0 = dP^\xi / d\lambda_1$ (densité de P^ξ pr à λ_1) et où $y = (y_1, ..., y_N)$.

Si F_0 est la **fonction de répartition** associée à P^ξ et si l'on pose $Z_n = F_0(Y_n)$ et $Z = (Z_1 ,..., Z_N)$, alors la densité de probabilité h de P^Z pr à λ_N ne dépend pas de P :

(6) $\quad h(z) = N ! . \mathbf{1}_C ,$

où $C = \{z \in]0, 1[^N : 0 < z_1 < ... < z_N < 1\}$.

La propriété (6) est très utilisée en **Statistique non paramétrique**.

(iv) Si X est un échantillon iid selon une **loi discrète** P^ξ (de densité f_0 pr à une **mesure de comptage**), il peut exister des valeurs multiples X_n dans X. On peut alors écrire :

(7)
$$P([X^{(1)} = ... = X^{(n(1))} = x_1] \cap ... \cap [X^{(n(1)+...+n(M-1)+1)} = ... = X^{(n(1)+...+n(M))} = x_M])$$
$$= (n_1 ! ... n_M !)^{-1} . n ! . \Pi_{m=1}^{M} \{f_0(x_m)\}^{n(m)},$$

où l'on note $n = \Sigma_{m=1}^{M} n_m$ et n(m) pour n_m .

(v) Si X est iid selon P^ξ (dont la fr est notée F_0), alors la coordonnée $X^{(n)}$ admet, $\forall$ n $\in \{1 ,..., N\}$, pour fr propre (ie pour **fonction de répartition marginale**) :

(8) $\quad G_{(n)}(x) = P(X^{(n)} \leq x) = n . C_N^n . \int_0^{F0(x)} u^{n-1} (1 - u)^{N-n} du.$

Si, de plus, F_0 est dérivable, de dérivée (densité) f_0 , alors la lp de $X^{(n)}$ admet pour densité :

(9) $\quad g_{(n)}(x) = n . C_N^n . (F_0(x))^{n-1} . \{1 - F_0(x)\}^{N-n} . f(x).$

La formule (8) supposait P^ξ absolument continue pr à λ_1 . Dans le cas discret (absolue continuité « discrète »), on obtient :

(8)' $\quad G_{(n)}(x) = \Sigma_{\alpha=n}^{N} C_N^\alpha (F_0(x))^\alpha . \{1 - F_0(x)\}^{N-\alpha}.$

(vi) Si $R = (R_1 ,.., R_N)$ est la **statistique de rang** associée à X, on a (par définition) :

(10) $\quad X^{(n)} = X_{\sigma(n)}$ (ou $= X_p$) $\Leftrightarrow$ $n = R_{\sigma(n)}$ (ou $= R_p$).

Autrement dit, $X^{(R(n))} = X_n$, $\forall$ n $\in \{1 ,..., N\}$ (en notant R(n) pour désigner R_n).

(vii) Une statistique d'ordre vérifie des inégalités classiques, eg :

(a) les **inégalités de G. STYAN - H. WOLKOWICZ**, ie, $\forall$ n $\in \{1 ,..., N\}$:

(11)
$$X^{(1)} \leq \bar{X}_N - (N - 1)^{-1/2} S_N ,$$
$$\bar{X}_N - \{(N - n) / n\}^{1/2} . S_N \leq X^{(n)} \leq \bar{X}_N + \{(N - 1) / (N-n+1)\}^{1/2} . S_N ,$$
$$X^{(N)} \geq \bar{X}_N + (N - 1)^{-1/2} S_N ,$$

expressions dans lesquelles $\bar{X}_N$ est la **moyenne empirique** de l'échantillon X et $S_N^2 = N^{-1} \Sigma_{n=1}^N (X_n - \bar{X}_N)^2 = N^{-1} X' P X$ sa **variance** empirique (non corrigée) ;

(b) l'**inégalité de R.A. GROENEVELD**. Si $X_n \geq 0$, $\forall\, n \in \{1, ..., N\}$, on a :

(12) $0 \leq X^{(n)} \leq (N - n - 1)^{-1} N \,.\, \bar{X}_N$, $\quad \forall\, n \in \{1, ..., N\}$.

(viii) En Statistique non paramétrique, on utilise souvent la statistique d'ordre $X^{(.)}$ pour définir des statistiques linéaires de la forme (cf **statistique linéaire de rang**) :

(13) $L_N = N^{-1} \Sigma_{n=1}^N s(n / (N+1)) \,.\, h(X^{(n)})$,

dans laquelle la fonction $s : [0, 1] \mapsto \mathbf{R}$ est appelée (fonction) « score » (cf **fonction score**) et la fonction $h : \mathbf{R} \mapsto \mathbf{R}$ est une **fonction numérique** mesurable quelconque.