TEST DE LA MÉDIANE (F3, I2)
(24 / 09 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le **test de la médiane** est un **test d'homogénéité** fondé sur une **statistique linéaire de rang** de la forme :

(1) $S_N = \Sigma_{n=1}^{N} c_n . a_N (R_n)$.

(i) Ce test s'associe notamment au problème à deux **échantillons** (cf **problème à plusieurs échantillons**, **score**). Dans ce cadre, on pose, dans (1) :

(2) $c_n = 1, \quad n = 1 ,..., N,$

et :

(3) $a_N (n) = \begin{cases} 0 \text{ ssi } n \leq (N + 1)/2, \\ 1 \text{ ssi } n > (N + 1)/2, \end{cases}$

et l'on définit alors la **statistique de la médiane** suivante :

(4) $S_N = \Sigma_{n=1}^{N(1)} a_N (R_n)$,

dans laquelle $N = N_1 + N_2$ est la taille de l'échantillon d'ensemble (empilé) X, N_1 la taille du premier échantillon (aussi noté N(1)), et R_n le rang de la n-ième coordonnée de X (cf **statistique de rang**).

La statistique de la médiane (4) s'interprète donc comme le nombre d'observations du premier échantillon X, situées à droite de la **médiane** d'ensemble.

(ii) Sous l'hypothèse H_0 d'**homogénéité** (stricte) des deux échantillons (ie $P^{\xi(2)} = P^{\xi(1)}$), en notant ξ(i) pour désigner ξ_i), ses deux premiers **moments** sont donnés par :

(5) $E_0 S_N = \begin{cases} N_2 / 2 & \text{ssi } N \in 2\, \mathbf{N}^*, \\ N_2 (N - 1) / (2 N) & \text{ssi } N \in 2\, \mathbf{N} + 1, \end{cases}$

et :

(6) $V_0 S_N = \begin{cases} (N_1 N_2) / \{4 (N - 1)\} & \text{ssi } N \in 2\, \mathbf{N}^*, \\ (N_1 N_2) (N + 1)) / (4 N^2) & \text{ssi } N \in 2\, \mathbf{N}^*, \end{cases}$

et S_N suit la **loi hypergéométrique** :

$$\text{(7)} \quad S_N \sim^{H0} \begin{cases} \mathscr{H}\{N, N_2, 1/2\} & \text{ssi } N \in 2\,\mathbf{N}^*, \\ \mathscr{H}\{N, N_2, (N-1)/(2N)\} & \text{ssi } N \in 2\,\mathbf{N} + 1. \end{cases}$$

Les tests qui en résultent sont ainsi appelés **tests de la médiane**.

(iii) De tels tests sont adaptés aux **distributions** à « **queue** épaisse » (eg la première des **lois de LAPLACE** ou la **loi de CAUCHY**).

(iv) Plus généralement, dans le cadre du problème à $k \geq 2$ échantillons, on considère l'échantillon d'ensemble (échantillon « empilé ») X, on définit T_i comme le nombre d'observations de X situées à droite de la médiane de X, et l'on pose :

$$\text{(8)} \quad S_{N(1)\ldots N(k)} = 4 \,.\, \{\Sigma_{i=1}^{k} N_i^{-1} \; (T_{N(i)})^2\} - N,$$

où N(i) dénote, par commodité, la taille N_i de l'échantillon n° i.

Lorsque la taille des échantillons est grande (ie lorsque $N^* = \min_{i=1}^{k} N_i >> 0$), le **test de la médiane** utilise la **propriété asymptotique** suivante, valable sous certaines conditions :

(9) $\mathscr{L}(S_{N(1)\ldots N(k)}) \to_{N^* \to +\infty} \mathscr{X}_{k-1}^{2}$ (**loi du chi-deux** à k **degrés de liberté**).

Ceci permet de tester l'hypothèse d'homogénéité stricte :

$$\text{(10)} \quad H_0 : F_1 = \ldots = F_k$$

contre une alternative tq (**alternative de position**) :

$$\text{(11)} \quad H_1 : F_i\,(x_{i,n(i)}) = F\,(x_{i,n(i)} - \mu_i),$$

avec $\mu = (\mu_1, \ldots, \mu_k) = 0$ (cf **test de KRUSKAL-WALLIS**) et où i,n(i) dénote, par commodité, l'indice double (i, n_i) $(i = 1, \ldots, k)$.